Identifying adolescents at risk for depression: a prediction score performance in cohorts based in three different continents

Running title: Identifying adolescents at risk for depression

Thiago Botter-Maio Rocha[1,2], MD PhD; Helen L Fisher[3], PhD; Arthur Caye[2], MD PhD; Luciana Anselmi[4], PhD; Louise Arseneault[3], PhD; Fernando C. Barros[4], MD PhD; Avshalom Caspi[3,5], PhD; Andrea Danese[3,6,7], MD PhD; Helen Gonçalves[4], PhD; HonaLee Harrington[5], BA; Renate Houts[5], PhD; Ana M. B. Menezes[4], MD PhD; Terrie E Moffitt[3,5], PhD; Valeria Mondelli[8,9], MD PhD; Richie Poulton[10], PhD; Luis Augusto Rohde[1,2,11], MD PhD; Fernando Wehrmeister[4], PhD; Christian Kieling[1,2], MD PhD

1. Division of Child & Adolescent Psychiatry, Hospital de Clínicas de Porto Alegre, Brazil.
2. Department of Psychiatry, School of Medicine, Universidade Federal do Rio Grande do Sul, Brazil.
3. King's College London, Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, London, United Kingdom.
4. Post-Graduate Program in Epidemiology, Universidade Federal de Pelotas, Pelotas, Brazil.
5. Department of Psychology and Neuroscience, Duke University, Durham, North Carolina, United States.
6. King's College London, Department of Child & Adolescent Psychiatry, Institute of Psychiatry, Psychology & Neuroscience, London, UK.
7. National and Specialist CAMHS Clinic for Trauma, Anxiety, and Depression, South London and Maudsley NHS Foundation Trust, London, UK.
8. National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre, South London and Maudsley NHS Foundation Trust, King's College London, London, UK.
9. King's College London, Department of Psychological Medicine, Institute of Psychiatry, Psychology & Neuroscience, London, UK.
10. Dunedin Multidisciplinary Health and Development Research Unit, Department of Psychology, University of Otago, Dunedin, New Zealand.
11. National Institute of Developmental Psychiatry for Children and Adolescents, São Paulo, Brazil.

Corresponding author: Christian Kieling, Department of Psychiatry, School of Medicine, Universidade Federal do Rio Grande do Sul; Child & Adolescent Psychiatry Division, Hospital de Clínicas de Porto Alegre, Rua Ramiro Barcelos 2350 – 400N, Porto Alegre, 90035-003, RS, Brazil (ckieling@ufrgs.br)

Acknowledgements: Drs. Rocha, Fisher, Caye, Anselmi, Arseneault, Barros, Caspi, Danese, Gonçalves, Harrington, Houts, Menezes, Moffitt, Poulton, Wehrmeister and Kieling report no competing interests. Dr. Mondelli has received research funding from Johnson & Johnson, a pharmaceutical company interested in the development of anti-inflammatory strategies for depression,

but the research described in this paper is unrelated to this funding. Dr. Rohde has been on the speakers' bureau/advisory board and/or has acted as a consultant for Eli-Lilly, Janssen-Cilag, Novartis and Shire in the last three years. He receives authorship royalties from Oxford Press and ArtMed. He also received travel awards for taking part of 2014 APA meeting from Shire. The ADHD and Juvenile Bipolar Disorder Outpatient Programs chaired by him received unrestricted educational and research support from the following pharmaceutical companies in the last three years: Eli-Lilly, Janssen-Cilag, Novartis, and Shire.

This work is supported by research grants from Brazilian public funding agencies to Christian Kieling and Luis A. Rohde: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS). This article is based on data from the study "Pelotas Birth Cohort, 1993" conducted by Postgraduate Program in Epidemiology at Universidade Federal de Pelotas, currently supported by the Wellcome Trust through the program entitled Major Awards for Latin America on Health Consequences of Population Change. The E-Risk Study is funded by the UK Medical Research Council (G1002190). Additional support was provided by the National Institute of Child Health and Human Development (HD077482) and by the Jacobs Foundation. Louise Arseneault is the Mental Health Leadership Fellow for the UK Economic and Social Research Council. The Dunedin Study is supported by the New Zealand Health Research Council, New Zealand Ministry of Business, Innovation, and Employment, National Institute on Aging Grant R01AG032282, and UK Medical Research Council Grant MR/P005918/1. The Identifying Depression Early in Adolescence (IDEA) project is funded by an MQ Brighter Futures grant (MQBF/1 IDEA). Additional support was provided by the UK Medical Research Council (MC_PC_MR/R019460/1) and the Academy of Medical Sciences (GCRFNG\100281) under the Global Challenges Research Fund. The views expressed are those of the authors. None of the funders played any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

We are extremely grateful to the individuals who participated in the studies in each of the sites and to all members of the IDEA consortium and the study teams for their dedication, hard work, and insights. The authors thank all the ProDIA group for their assistance in the development of this work. The authors would like to especially thank Dr. João Ricardo Sato, PhD, from Universidade Federal do ABC, for his thoughtful insights into the initial version of this study and Dr. Rachel Latham, PhD, from King's College London, for assistance with checking the statistical analysis for the E-Risk study.

Drs Kieling and Rohde conceptualized the study. Drs Fisher, Anselmi, Arseneault, Barros, Caspi, Danese, Gonçalves, Harrington, Houts, Menezes, Moffitt, Poulton, Rohde, Wehrmeister and Kieling contributed to the study design and/or data collection. Drs Rocha, Fisher, Caye, Harrington, Houts and Kieling contributed to data analysis. Drs Rocha, Fisher, Caye, Arseneault, Caspi, Menezes, Moffitt, Mondelli, Rohde and Kieling contributed to data interpretation. Drs Rocha and Kieling contributed to the writing of the manuscript. Drs. Rocha and Kieling had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors were responsible for critical review of the manuscript for important intellectual content and all authors reviewed and approved the final version of the manuscript.

Abstract: 267 words
Main text: 3418 words
Figures and tables: 3 tables and 3 figures
Supplement: Yes
Key words: Adolescent; Depression; Cohort studies; Risk assessment; Prognosis

Facebook:

Study using @coorte1993 data from Brazil developed a risk score to predict the occurrence of #depression at the age of 18 in healthy 15 years-olds using only sociodemographic variables. The performance of the score was also assessed in #adolescents from United Kingdom and New Zealand, providing relevant insights for stratifying adolescents at risk for #depression <link to article placeholder>

Twitter:

In a new study @JAACAP authors developed a risk calculator to predict #depression in late #adolescence using only sociodemographic variables. @idea_mq @depadol <link to article placeholder>

Lay Summary

Using data from a prospective study from Brazil, the authors developed a score identify, among 15-year old healthy adolescents, who was at risk for developing depression at age 18 yers. The model include only sociodemographic variables easily obtainable directly from the adolescents: biological sex, skin color, drug use, school failure, social isolation, fight involvement, poor relationship with mother, poor relationship with father, poor relationship between parents, childhood maltreatment, history of running away from home. The model was also informative when assessed in two other independent studies, one from New Zealand and the other from the United Kingdom. These results suggest that the tool can be a promissing aid in identifying adolescents at high and at low risk for developing depression.

Abstract

Objective: Prediction models have become frequent in the medical literature, but most published studies are conducted in a single setting. Heterogeneity between development and validation samples has been posited as a major obstacle for the generalization of models. We aimed to develop a multivariable prognostic model using sociodemographic variables easily obtainable from adolescents at age 15 to predict a depressive disorder diagnosis at age 18, and to evaluate its generalizability in two samples from diverse socioeconomic and cultural settings.

Methods: Data from the 1993 Pelotas Birth Cohort were used to develop the prediction model, and its generalizability was evaluated in two representative cohort studies: the Environmental Risk (E-Risk) Longitudinal Twin Study and the Dunedin Multidisciplinary Health and Development Study.

Results: At age 15, 2,192 adolescents with no evidence of current or previous depression were included (44.6% males). The apparent C-statistic of the models derived in Pelotas ranged from 0.76 to 0.79, and the model obtained from a penalized logistic regression was selected for subsequent external evaluation. Major discrepancies between the samples were identified, impacting the external prognostic performance of the model (Dunedin and E-Risk C-statistics of 0.63 and 0.59, respectively). The implementation of recommended strategies to account for this heterogeneity among samples improved the model's calibration in both samples.

Conclusion: An adolescent depression risk score comprising easily obtainable predictors was developed with good prognostic performance in a Brazilian sample. Heterogeneity among settings was not trivial, but strategies to deal with sample diversity were identified as pivotal for providing better risk stratification across

samples. Future efforts should focus on developing better methodological approaches for incorporating heterogeneity in prognostic research.

Introduction

The field of prognostic research has seen a substantial rise in publications of prediction modeling studies in the last decade.[1] This increase prompted significant advances in several medical specialties.[2,3] However, most published prognostic models have been assessed in a single setting.[4,5] Performance results obtained from model-development studies are frequently not achieved in validation trials, when evaluated. This inconsistency can be explained either by an overoptimistic prognostic performance from an overfitted model or by significant discrepancies between development and validation samples.[6]

When assessing external validation across datasets, heterogeneity among prognostic studies is the norm rather than the exception.[7] Differences in assessment strategies, frequency of outcome and/or studied factors, or availability of variables of interest could impose considerable difficulties for comparison purposes, impairing model generalizability. Current methodological guidelines recommend a set of careful development steps from derivation to external validation and ultimately use in clinical practice.[8] In this process, understanding the similarities and differences between samples is essential,[9] as guidelines suggest that a model with poor external performance should be updated before being discarded.[6,10] This procedure integrates information obtained from new data to the developed model, potentially improving its prognostic ability.[4,11] Even consolidated prediction models, such as the Framingham score for cardiovascular outcomes, face important drawbacks when applied in samples somewhat diverse from the original,[12] demanding model adjustments to enhance generalizability to different settings.[4,6]

Up to now, the majority of psychiatric composite prognostic models studies have focused on model development, with very few being adequately validated in independent samples.[13-15] In contrast to other areas of medicine, where hard outcomes are more easily defined, imprecise characterization of psychiatric outcomes imposes additional barriers for accurate prognostic model development and validation, as reliability of common mental disorders such as depression has been shown to be low.[16] Substantial heterogeneity in clinical presentation and high rate of comorbidity produce additional obstacles for prediction of psychiatric disorders, as different assessment strategies influence the likelihood of endorsing a diagnosis.[17]

Prediction of psychosis, the most prolific and consolidated area in prognostic psychiatry, has greatly advanced at group level. However, it still faces challenges in prediction at the individual subject level.[18] Prediction of major depressive disorder (MDD), the leading cause of mental health-related disease burden globally, is still in its infancy, relying mainly on single predictors for definition of at-risk individuals, with only a few studies combining risk factors.[19] Following recently published standards for appropriate development and validation of psychiatric prediction models,[20] using the most recent methodological recommendations,[1,6] and state-of-the-art statistical strategies,[21,22] the present study aims to derive and evaluate the generalizability of a psychiatric prediction model across samples from different sociocultural backgrounds.

Using data obtained from globally-relevant longitudinal population-based cohorts, our first goal was to develop a multivariable prognostic model to evaluate the risk of

developing a depressive episode by late adolescence in a Brazilian sample of adolescents with no evidence of previous depression, using *a priori* selected, easily-obtainable sociodemographic variables, collected directly from adolescents. Our second aim was to evaluate the impact of heterogeneity on its generalization to two diverse sociocultural contexts, as well as to assess strategies to overcome these limitations.

Methods

Samples and participants

We derived our prediction model using data exclusively from the largest cohort available, the 1993 Pelotas Birth Cohort, a prospective study set in Brazil, and then evaluated the generalizability of findings in two diverse samples: the Environmental Risk (E-Risk) Longitudinal Twin Study, from the UK, and the Dunedin Multidisciplinary Health and Development Study, from New Zealand. Details about the three cohorts are reported elsewhere,[23-25] and in the online supplementary material. In brief, in the Pelotas study, all 5,249 children born live in the city of Pelotas in 1993 were enrolled in the study. The original goals of the 1993 Cohort were to evaluate trends in maternal and child health indicators to assess associations between early life variables and later outcomes. At the wave for ages 18-19 years old, the retention rate was 81.3% of the original sample. The Environmental Risk (E-Risk) Longitudinal Twin study tracks the development of a nationally-representative birth cohort of 2,232 British twin children born in England and Wales in 1994-1995.[20] The sample was constructed in 1999-2000, when 1,116 families with same-sex 5-year-old twins (93% of those eligible) participated in home-visit assessments. The Dunedin

Study is a longitudinal investigation of health and behavior in a complete birth cohort. All study participants (N=1,037; 91% of eligible births; 52% male) were born between April 1972 and March 1973 in Dunedin, New Zealand.

To be included in the final analysis, an evaluation for a depressive episode in late adolescence (18-19 years) was required. Exclusionary criteria were applied, filtering out those with intelligence quotient <70, and/or no signs of puberty by 15 years of age. Additionally, as our intention was to provide an alternative risk screening strategy beyond using previous depressive episodes or sub-threshold depressive symptoms, individuals with any suggestive evidence of a current or previous MDD diagnosis by the age of risk ascertainment were excluded from the final sample (see online supplement). As the E-Risk sample was not evaluated at age 15, we have selected the most comparable assessment wave, namely age 12. Given the age difference at baseline between the E-Risk sample and the other samples, puberty was not considered an exclusionary criterion for this sample.

Assessment and definition of predictor variables

Selection of predictors was based on scientific literature review and authors' clinical expertise,[26] but constrained to their availability in the Pelotas dataset. As we aimed for real-word implementation, following a pragmatic approach,[27] we included variables readily available, not too costly to obtain, and simple to evaluate.[20,22] We adopted an *a priori* defined criterion to use only variables directly obtained from the adolescents in the Pelotas study at the 15-years' assessment wave to mirror the reality in routine practice, selecting 11 variables related to inherent characteristics (biological sex, skin color); problematic behavior indicators (drug use, school failure, social isolation, fight

involvement); and markers of household dysfunction (poor relationship with mother, poor relationship with father, poor relationship between parents, childhood maltreatment, ran away from home). For comparison purposes, the harmonization of selected variables among cohorts was performed *a priori* by consensus among investigators from each site. Further details on variables' assessment strategies are provided in Table S1, available online.

Assessment and definition of the outcome variable

In each sample, the outcome of interest was a categorical diagnosis of depression in late adolescence. In the Pelotas cohort, trained psychologists interviewed the participants at ages 18-19 years in 2011-12 with a structured interview for current major depressive disorder diagnosis using the Mini-International Neuropsychiatric Interview (MINI) – DSM-IV-TR criteria, MDD section, assessing symptoms in the previous 2 weeks. For the E-Risk sample, a major depressive disorder diagnosis in the previous 12 months was assessed using the Diagnostic Interview Schedule (DIS) at age 18 based on DSM-IV criteria in 2012-14. In the Dunedin cohort, past-year major depressive disorder diagnosis was evaluated using the DIS at age 18 following DSM-III-R criteria in 1990-91.

Statistical analysis

A detailed description of statistical procedures used can be found in Supplement 2. In an effort to enhance our model's reproducibility, we transparently described the process of model development and validation. Using data from the Pelotas cohort, we developed a baseline model using binary logistic regression (LR) analysis – the most common statistical strategy in prognostic research. As overfitting is a major reason for

irreproducibility, we derived six new models from the same dataset introducing different strategies of model penalization – one penalized LR model using penalized maximum likelihood estimation (PMLE) and five models with increasing degrees of penalization using the Elastic-Net machine learning algorithm.[21] Comparing penalized models' parameters to our baseline model, we selected for validation the one with more balanced performance measures.

To evaluate the performance of the selected model in new observations, we first internally validated it using standard bootstrapping procedures to measure undue optimism in the model's performance metrics, which happens when the model is evaluated directly in the derivation cohort (apparent performance). Second, we quantified the model's prognostic performance in independent observations in two prospective cohorts from diverse contexts.

When assessing a given model's prediction in independent samples, its performance may be influenced by differences between derivation and validation cohorts.[6] Differences can be related not only to distribution of participant characteristics (case-mix), but also from true differences in predictor effects. To take this into account, we adopted a sequence of recommended approaches.[6,22] We calculated a case-mix-corrected and a refitted model for each sample, and the obtained metrics were used as performance parameters for each sample. Additionally, some of the originally selected variables were not available in all the cohorts, a likely situation in real-world model application. Instead of excluding these variables, we evaluated the amount of the original model's information lost by this mismatch.[21] Finally, we evaluated the impact of between-study heterogeneity by aggregating all cohorts into an overall

sample to model cohort differences either in baseline risk or in predictor effects (see Supplement 3).[28]

All statistical analyses were performed using R software, version 3.4.4. A complete-case analysis strategy was used, excluding participants with any missing data. A multiple imputation procedure using R package *mice* was applied to assess missing data impact (see Table S2 and Figure S1).

Results

Sample characteristics

A flowchart for each cohort can be seen in Figure 1a-c. From the original sample size of 5,249 individuals in the Pelotas cohort, 81.3% were retained up to the 18-19 years assessment, and 2,192 were included for final analyses after applying exclusion criteria. For E-Risk and Dunedin samples, from the 2,232 and 1,037 initially assessed individuals, 1,144 (51.3%) and 739 (71.3%) were available for assessment after exclusion criteria were applied, respectively. Comparisons on key characteristics between retained and excluded samples for the Pelotas cohort are provided in Table S3.

Table 1 displays descriptives for both depression outcome and selected predictors in each sample. Noteworthy disparities were identified regarding rates of school failure, social isolation, fight involvement, and running away. Additionally, family relationships were not assessed in the E-Risk Study. The MDD prevalence in Pelotas, E-Risk and Dunedin samples was 3.1%, 17.7% and 16.8%, respectively. Differences

in outcome prevalence among cohorts may have reflected differences in timeframe for outcome assessment (2 weeks vs 12 months).

Model development and validation

Performance measures showed better results for models using LR strategies compared to machine learning Elastic-Net approaches. In the Pelotas sample, discriminative capacity to parse between adolescents who later developed depression at age 18 and those who did not, assessed by the C-statistic, ranged from 0.76 to 0.79, indicating overall good discrimination, as shown in Table 2.

Predictably, the baseline model showed the best combination of performance metrics. Among penalized models, the PMLE model presented better performance when compared to all Elastic-Net models. As non-penalized models face a greater risk of overfitting, we proceeded to the next step with both LR models for comparison. We internally validated each using bootstrapping evaluation with 1,000 iterations. As expected, measurement of optimism – difference between apparent and bias-corrected performance metrics – was lower for the PMLE when compared to the LR model (ΔC-statistic: 0.067 vs 0.098; ΔSlope: -0.004 vs 0.548; ΔR^2: 0.034 vs 0.149, respectively), suggesting lower overfitting and higher probability of reliable results when applied to independent samples. Additionally, as shown in Figures S2a-b, the PMLE model was also more calibrated, with a 60% reduction in mean square error compared to the LR model. Therefore, the PMLE model was selected as the Pelotas final model, with a C-statistic of 0.78 (Bootstrap-corrected 95%CI: 0.73 to 0.82).

Using the most common external validation strategy, the linear predictor derived from the selected Pelotas model (see Table S4) was applied to the other samples. There was an expected drop of the performance metrics in both independent cohorts (E-Risk: C-statistic=0.59 – Bootstrap-corrected 95%CI: 0.55 to 0.63; Dunedin: C-Statistic=0.63 – Bootstrap-corrected 95%CI: 0.59 to 0.67). The performance results for each step of the validation process can be seen in Table 3.

Model updating

As variables from both independent datasets did not perfectly pair with the set selected from the Pelotas study, we calculated the amount of information lost due to this mismatch.[21] In the E-Risk dataset, 13.1% of original model information was unavailable, mainly from the household dysfunction indicators. In Dunedin, this percentage was lower, at around 6.9%.

Considering the relevant heterogeneity among cohorts, we evaluated if the integration of information from the external cohorts could produce improvement in model performance, in line with current methodological recommendations.[4] As differences in outcome prevalence were not trivial, we updated the Pelotas model by correcting its intercept for each cohort. In both validation samples, the updated model produced better calibration, reducing all measures of calibration error (see Supplement 2 and Figure S3a-d).

Exploratory analyses

The merger of all three cohorts into an aggregated sample to assess between-cohorts heterogeneity increased the total number of individuals to 4,075, of which 395 (9.7%)

presented a positive outcome. Given the majority of individuals were from the Pelotas cohort (53.8%), the C-statistic was also 0.78 (Bootstrap-corrected 95%CI: 0.75 to 0.80), but showed lower overfitting after internal validation using bootstrapping (see Figure 2a-b). Inclusion of each cohort's main-effects and their interaction terms with all predictors into a PMLE model suggested that not only disparities in case-mix, as shown in Table 1, but also between-cohorts differences in predictor effects might have influenced external validation results, particularly considering the difference in the ran-away and fight involvement variables (see Figure 3).

Discussion

Following current standards for psychiatric prognostic research,[20] our study proposes a multivariable model developed in a Brazilian cohort to predict, among adolescents with no evidence of previous depression, the risk of developing a depressive episode in late adolescence. Our model showed beyond chance results of discrimination and calibration, with metrics comparable to established prognostic models from other areas of medicine,[3,29] and could be viewed as a promising aid to adolescent depression risk stratification.[30]

Evaluation in independent samples is deemed essential for generalization of findings. Disparities among samples are frequently seen as major obstacles for model validation, replication and generalizability. However, as the *Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis* (TRIPOD) statement emphasizes, the term validation can be misleading, recommending that an external validation should quantify the model's prognostic performance in a new sample, not simply classifying it as a positive or negative validation.[4,31,32] This

broader validation approach promotes not only the assessment of the model's performance in the new sample, but also facilitates understanding of why the results differ.

For this study, we assessed the validation performance of the model developed in our Brazilian sample in two population-based longitudinal cohorts from two different continents. The development of a model in one middle-income country and its external validation in samples representing diverse sociocultural and economic contexts, using different assessment strategies for data collection at different time periods among them, may help evaluate if and where its results can be generalized. Our results suggest that, albeit adaptations should be applied on the original model to enhance external clinical utility, the original prognostic model could be applied in multiple other contexts despite major differences in assessment strategies, socioeconomic characteristics and cultural influences. Given such profound differences, it was expected that the developed model could not be easily transported to new settings.[9] Even though lower in degree, our model kept a valid and beyond chance prognostic capacity in parsing future risk of depression among the adolescents in the independent cohorts, especially when heterogeneity among samples was accounted for (see online Supplement).

Early identification of individuals at higher risk for psychiatric disorders could potentially lessen the massive burden imposed by these conditions. Positive family history of depression and the presence of sub-threshold depressive symptoms have been the most commonly used criteria for identifying at-risk individuals.[33] Although these strategies have been replicated, reliance on single predictors restricts their

prognostic contribution, not accounting for a wider range of risk. Additionally, from a pragmatic perspective, the need of trained staff for proper evaluation of such predictors limits their potential implementation, given that access to treatment has been systematically highlighted as a major barrier for child and adolescent mental health care.[34]

Our study has several strengths. We developed a prognostic model for MDD according to most recent guidelines in prognostic research and transparent reporting,[6,20] using modern, state-of-the-art statistical strategies,[21,22] with broad external validation assessment. Comprising only 11 predictors, all easily-obtainable, quick to assess, and collected directly from the adolescent, with no need for highly specialized training, external informants or laboratory analyses, our results could be seen as promising if further replicated. Additionally, consistent with the evidence-based pragmatic psychiatry initiative,[27] we opted to prioritize simplicity over accuracy, selecting predictors that could be more easily and broadly implemented, enhancing probability of future clinical use and patient acceptance.

Significant limitations of our study also need to be considered. Having based the development of our prognostic model on the Pelotas cohort, an ongoing study not primarily focused on mental health, availability of variables of interest was limited to those previously collected, precluding the use of some potentially relevant factors. MDD diagnosis was assessed at the 18-19 years wave by evaluating symptoms in the two weeks before the interview, limiting comparability to other epidemiological cohort studies, as well as reducing the prevalence of the outcome of interest. Consequently, the number of outcome events per selected variable was lower in the

Pelotas sample (EPV=6.27), increasing the risk of overfitting.[20-22] Strategies such as machine learning regularization methods, with shrinkage and selection of predictors, as well as measurement of performance optimism, were implemented to constrain the impact of this limitation. The proposed model is also not necessarily prognostic of earlier or later onsets of depression. Furthermore, having that we are analyzing individuals at higher risk of a major depression diagnosis, we could not discard the chance that all self-report assessments were biased by this risk. Additionally, as our goal was to provide a risk stratification tool that could be supplementary to current strategies of risk evaluation, we opted to exclude individuals with any evidence of previous or current depressive episodes, as the occurrence of a depressive episode already heightens the risk of subsequent depression. As this strategy resulted in a significant number of exclusions that could have biased our findings, we compared the covariates between included and excluded samples (Table S3), with anticipated differences between them, and performed sensitivity analyses (see Table S6 and Figure S4), in which similar performance results were identified.

Another relevant shortcoming is the differences in predictors' availability and assessment strategies among cohorts, which could have influenced results obtained in the external validations. The unavailability of assessment data at the age 15 in the E-Risk sample could have impacted the comparability among the samples, as puberty is a well-know risk contributor for depression,[36] and could therefore have contributed to the performance result of the model in that sample. *A priori* harmonization of variables and measurement of information lost due to mismatching variables were applied to minimize the effect of these limitations. Also, we were constrained to variables assessed in each cohort study, which precluded important predictors being

included in our model, and the included variables could be carrying prognostic information from uncollected predictors, which could have contributed to predictor effects' discrepancies shown in Figure 3. Finally, we could not evaluate, in the present study, the potential impact of the developed model on clinical decision-making.[20]

Exploratory analyses suggested that information generated by our model increased prognostic ability above and beyond established risk factors such as subsyndromal symptoms and a positive family history of depression (Table S7). At the same time, the risk score was also associated, to a lesser degree, to other diagnostic outcomes (C-statistic range: 0.64 to 0.70) (Table S8). In line with the current literature on the early detection of psychopathology in youth,[37] we believe that a transdiagnostic approach could be considered, despite its limitations,[38] as psychiatric prognostic models' specificity is likely to be low and as less specific preventive interventions could promote meaningful changes in psychiatric burden, either from individual or public health perspectives.[9,39]

We present the development of a prognostic model for major depressive disorder among Brazilian adolescents, externally evaluated in two samples from diverse sociocultural contexts using different strategies for data collection than the original cohort. Heterogeneity among studies was high and possibly accounted for major discrepancies in prognostic performance, probably related not only to different case-mix but also coefficients' weights.[6] Future studies should pursue methodological strategies for embracing heterogeneity among samples, instead of avoiding it, thus

producing results which are more likely to be translated into clinical practice across a range of contexts.

Table 1. Sample description for each cohort[a]

	Pelotas (Brazil)	E-Risk (UK)	Dunedin (New Zealand)
Included sample	2,192	1,144	739
Assessment age	15 years	12 years	15 years
Male sex	977 (44.6%)[b]	520 (45.5%)[b]	375 (50.7%)[c]
White skin color	1,478 (67.4%)[b]	1,040 (90.9%)[c]	NA[e]
Childhood maltreatment			
None	1,539 (70.2%)[b]	963 (84.2%)[c]	489 (66.2%)[d]
Probable	390 (17.8%)	139 (12.2%)	187 (25.3%)
Severe	263 (12.0%)	42 (3.7%)	63 (8.5%)
School failure	1,127 (51.4%)[b]	212 (18.5%)[c]	80 (10.8%)[d]
Social isolation	231 (10.5%)[b]	63 (5.5%)[c]	70 (9.5%)[b]
Fights	211 (9.6%)[b]	130 (11.4%)[b]	12 (1.6%)[c]
Ran away from home	80 (3.6%)[b]	9 (0.8%)[c]	49 (6.6%)[d]
Any drug use	1,367 (62.4%)[b]	569 (49.7%)[c]	592 (80.1%)[d]
Relationship with mother		NA	
Great	1,417 (64.6%)		
Very good	430 (19.6%)		
Good	264 (12.0%)		
Regular	68 (3.1%)		
Bad	13 (0.6%)		
Relationship with father		NA	22.0 [5.4][f]
Great	1,019 (46.5%)		
Very good	434 (19.8%)		
Good	370 (16.9%)		
Regular	237 (10.8%)		
Bad	132 (6.0%)		
Relationship between parents		NA	
Great	886 (40.4%)[b]		345 (46.7%)[c]
Very good	421 (19.2%)		278 (37.6%)
Good	404 (18.4%)		91 (12.3%)
Regular	301 (13.7%)		23 (3.1%)
Bad	180 (8.2%)		2 (0.3%)
Depression prevalence	69 (3.1%)[b,g]	202 (17.7%)[c,h]	124 (16.8%)[d,h]

Results are shown as number of individuals (percentage) for categorical variables, and as mean [standard deviation] for continuous variables for individuals included in the final analyses. NA: Data not available in the cohort.

[a] See Table S1 for assessment strategies applied to each cohort.

Superscript letters “b”, “c” and “d” denote column differences among the samples: different letters show significant and equal letters indicate non-significant differences from each other, assessed by chi-square (χ^2) test at a 0.05 level. For variables with more than two categories, the superscript letters were placed in the first row of the variable and represent the assessment of the variable as a group, not per row.

[e] Skin color was not assessed in the cohort. Fewer than 7% of the cohort has any non-white ancestry.

[f] Parent Attachment Scale score (range: from -6 to 28) - Adolescent assessment about the relationship with both parents.

[g] Presence of symptoms reaching diagnostic criteria within a two-week period before assessment.

[h] Presence of symptoms reaching diagnostic criteria within a twelve-month period before assessment.

Table 2. Apparent performance parameters obtained from the models derived from the Pelotas' dataset

	Model Parameters						
	LR	PMLE[a]	Ridge[b]	0.25[b]	0.50[b]	0.75[b]	LASSO[b]
R^2	0.15	0.12	0.12	0.10	0.10	0.10	0.10
LR χ^2 [c]	81.90	66.17	63.30	54.40	54.32	54.71	54.10
Brier score[d]	2.88	2.93	2.93	2.95	2.95	2.95	2.95
C-statistic[e]	0.79	0.78	0.78	0.76	0.76	0.76	0.76
Calibration slope	1.00	1.26	1.35	1.47	1.42	1.38	1.39

Higher results for R^2, LR χ^2 and C-statistic, lower results for Brier score, and results closer to 1 for Calibration slope indicate better model performance.

0.25: Elastic-Net with alpha=0.25; 0.50: Elastic-Net with alpha=0.50; 0.75: Elastic-Net with alpha=0.75; LASSO: Least absolute shrinkage and selection operator; LR: Logistic regression; PMLE: Penalized maximum likelihood estimation; Ridge: Ridge regression.

R^2: Nagelkerke's R^2; LR χ^2: Likelihood Ratio χ^2; C-statistic: Concordance statistic, or area under the curve of the receiver operating characteristic (AUC-ROC); Brier score: Quadratic scoring rule that combines calibration and discrimination; Calibration slope: measure of agreement between observed and predicted risk of the event (outcome) across the whole range of predicted values.

[a] The penalty factor used in the PMLE was empirically obtained from our data.

[b] For the Elastic-Net approach, we have *a priori* defined a grid of values for the hyperparameter alpha, ranging from 0 (full Ridge) to 1 (full LASSO), with increments of 0.25. For each alpha value, a 10-fold cross-validation was used to select the penalty coefficient (lambda) that minimized the mean squared prediction error, which was then used for shrinkage of coefficients and/or variable selection. See Table S4 for model's coefficients.

[c] *All LR χ^2 p-values <0.001.*

[c] Multiplied by 10^2.

[d] The C-statistic ranges from 0.5 for non-informative models to 1.0 for perfect models.

Table 3. Comparative results for each step of model performance in the three cohorts

Performance parameter	Description	Pelotas		E-Risk			Dunedin		
		Apparent validation	Internal validation	External validation	Case-mix-corrected model[a]	Refitted model[b]	External validation	Case-mix-corrected model[a]	Refitted model[b]
C-Statistic	Concordance statistic, equal to the area under the curve of the receiver operating characteristic (AUC-ROC) in binary endpoints	0.78	0.71	0.59	0.66	0.62	0.63	0.68	0.67
Calibration-in-the-large	An overall measure of calibration, compares mean observed with mean predicted in the validation dataset	0.00	0.02	2.37	0.02	0.00	2.26	-0.06	0.00
Calibration slope	Measure of agreement between observed and predicted risk of the event (outcome) across the whole range of predicted values	1.26	1.00	0.58	0.99	1.20	0.77	0.98	1.24
R^2	Measure of overall goodness-of-fit of the model	0.12	0.06	0.03	0.04	0.05	0.05	0.05	0.09
Brier score	Quadratic scoring rule that combines calibration and discrimination	0.03	0.03	0.17	0.02	0.14	0.16	0.02	0.13
Emax	Maximum absolute error in predicted probabilities	0.19	0.03	0.29	0.01	0.09	0.38	0.01	0.11
Available information for the assessment of model performance		100%		86.9%			93.1%		

Higher results for C-statistic and R^2, lower results for Brier score and Emax, results closer to 0 for Calibration-in-the-large, and results closer to 1 for Calibration slope indicate better model performance.
[a] Reference values indicating the model's performance under the assumption that Pelotas model's coefficients are fully correct for the validation setting, simulating similar case-mix between samples.[22]
[b] Reference values indicating the model's performance after refitting predictors' coefficients that would be optimal for the validation sample.[22] (See eMethods 2 for further details.)

Figure legends

Figure 1a-c. Flowcharts for each included cohort study.

Figure 2. Performance measures of the aggregated sample model: a) the area under the receiver operating characteristic (ROC) curve and the bootstrapped 95% confidence interval (indicated by grey shading) of the C-statistic, and b) calibration plot after internal validation using 1,000 iterations bootstrapping. Apparent and bias-corrected results were plotted as a nonparametric calibration curve, estimated over a sequence of predicted values vs. observed values using a smoothing technique.

Figure 3. Comparison of the prognostic contribution of each included variable in each cohort to the aggregated sample prediction model of adolescent depression, stratified by sex for Brazil, UK, and New Zealand cohorts. Predictors' beta coefficients from penalized logistic regression are shown as bars in the x-axis. Positive values represent greater risk, and negative values represent lower risk of the outcome. The results shown are derived from values presented in Table S5. Some of the variables previously included in the Pelotas model were excluded for comparability among datasets.

REFERENCES

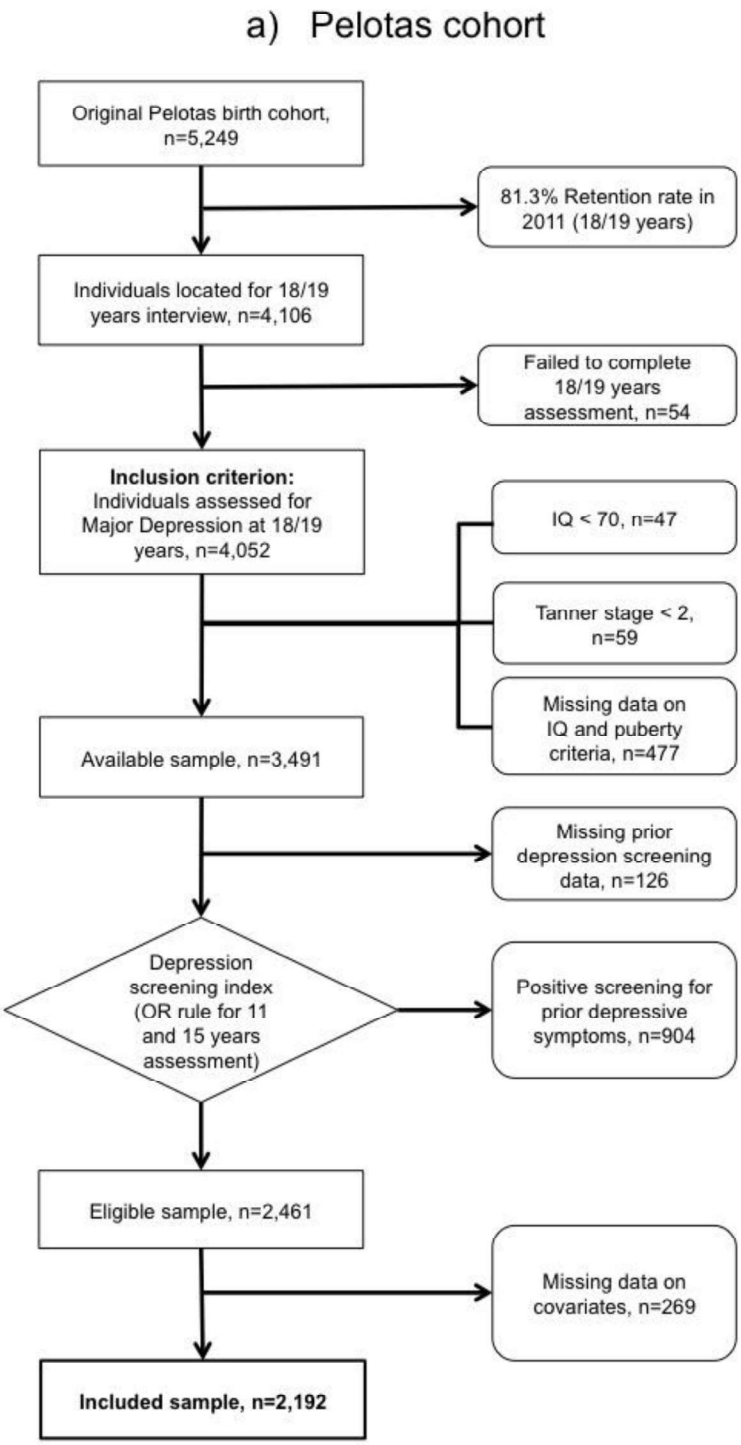
1. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med.* 2013;10(2):e1001381.
2. Kamath PS, Wiesner RH, Malinchoc M, et al. A model to predict survival in patients with end-stage liver disease. *Hepatology.* 2001;33(2):464-470.
3. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation.* 2008;117(6):743-753.
4. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart.* 2012;98(9):691-698.
5. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. *BMJ.* 2011;343:d7163.
6. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol.* 2015;68(3):279-289.
7. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med.* 2013;32(18):3158-3180.
8. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol.* 2015;68(1):25-34.
9. Fusar-Poli P, Werbeloff N, Rutigliano G, et al. Transdiagnostic Risk Calculator for the Automatic Detection of Individuals at Risk and the Prediction of Psychosis: Second Replication in an Independent National Health Service Trust. *Schizophr Bull.* 2019;45(3):562-570.
10. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ.* 2016;353:i3140.
11. Snell KI, Hua H, Debray TP, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol.* 2016;69:40-50.
12. D'Agostino RB, Grundy S, Sullivan LM, Wilson P, Group CRP. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA.* 2001;286(2):180-187.
13. Fusar-Poli P, Rutigliano G, Stahl D, et al. Development and Validation of a Clinically Based Risk Calculator for the Transdiagnostic Prediction of Psychosis. *Jama Psychiatry.* 2017;74(5):493-500.
14. Birmaher B, Merranko JA, Goldstein TR, et al. A Risk Calculator to Predict the Individual Risk of Conversion From Subthreshold Bipolar Symptoms to Bipolar Disorder I or II in Youth. *J Am Acad Child Adolesc Psychiatry.* 2018;57(10):755-763.e754.
15. Kessler RC, Warner CH, Ivany C, et al. Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army Study To Assess Risk and rEsilience in Servicemembers (Army STARRS). *JAMA Psychiatry.* 2015;72(1):49-57.
16. Regier DA, Narrow WE, Clarke DE, et al. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *Am J Psychiatry.* 2013;170(1):59-70.
17. Zimmerman M, Ellison W, Young D, Chelminski I, Dalrymple K. How many different ways do patients meet the diagnostic criteria for major depressive disorder? *Compr Psychiatry.* 2015;56:29-34.

18. Studerus E, Ramyead A, Riecher-Rössler A. Prediction of transition to psychosis in patients with a clinical high risk for psychosis: a systematic review of methodology and reporting. *Psychol Med.* 2017;47(7):1163-1178.
19. King M, Walker C, Levy G, et al. Development and Validation of an International Risk Prediction Algorithm for Episodes of Major Depression in General Practice Attendees The PredictD Study. *Archives of General Psychiatry.* 2008;65(12):1368-1376.
20. Fusar-Poli P, Hijazi Z, Stahl D, Steyerberg EW. The Science of Prognosis in Psychiatry: A Review. *JAMA Psychiatry.* 2018.
21. Harrell FE. *Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis.* New York: Springer; 2001.
22. Steyerberg EW. *Clinical prediction models : a practical approach to development, validation, and updating.* New York, NY: Springer; 2009.
23. Moffitt TE, Team E-RS. Teen-aged mothers in contemporary Britain. *J Child Psychol Psychiatry.* 2002;43(6):727-742.
24. Poulton R, Moffitt TE, Silva PA. The Dunedin Multidisciplinary Health and Development Study: overview of the first 40 years, with an eye to the future. *Soc Psychiatry Psychiatr Epidemiol.* 2015;50(5):679-693.
25. Victora CG, Hallal PC, Araújo CL, Menezes AM, Wells JC, Barros FC. Cohort profile: the 1993 Pelotas (Brazil) birth cohort study. *Int J Epidemiol.* 2008;37(4):704-709.
26. Newton S, Docter S, Reddin E, Merlin T, Hiller J. Depression in Adolescents and Young Adults: Evidence Review. Adelaide: Adelaide Health Technology Assessment (AHTA), University of Adelaide; 2010. url:https://www.adelaide.edu.au/ahta/pubs/depression-in-adolescents-and-young-adults.pdf. Accessed November, 3, 2018.
27. Paulus MP. Evidence-Based Pragmatic Psychiatry-A Call to Action. *JAMA Psychiatry.* 2017;74(12):1185-1186.
28. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol.* 2016;69:245-247.
29. Lambertini M, Pinto AC, Ameye L, et al. The prognostic performance of Adjuvant! Online and Nottingham Prognostic Index in young breast cancer patients. *Br J Cancer.* 2016;115(12):1471-1478.
30. Kieling C, Adewuya A, Fisher HL, et al. Identifying depression early in adolescence. *Lancet*. 2019;3:211-212.
31. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med.* 2015;162(1):55-63.
32. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-73.
33. Hetrick SE, Cox GR, Witt KG, Bir JJ, Merry SN. Cognitive behavioural therapy (CBT), third-wave CBT and interpersonal therapy (IPT) based interventions for preventing depression in children and adolescents. *Cochrane Database Syst Rev.* 2016(8):CD003380.
34. Costello EJ, He JP, Sampson NA, Kessler RC, Merikangas KR. Services for adolescents with psychiatric disorders: 12-month data from the National Comorbidity Survey-Adolescent. *Psychiatr Serv.* 2014;65(3):359-366.
35. Birmaher B, Williamson DE, Dahl RE, et al. Clinical presentation and course of depression in youth: does onset in childhood differ from onset in adolescence? *J Am Acad Child Adolesc Psychiatry.* 2004;43(1):63-70.
36. Thapar A, Collishaw S, Pine DS, Thapar AK. Depression in adolescence. *Lancet.* 2012;379(9820):1056-1067.

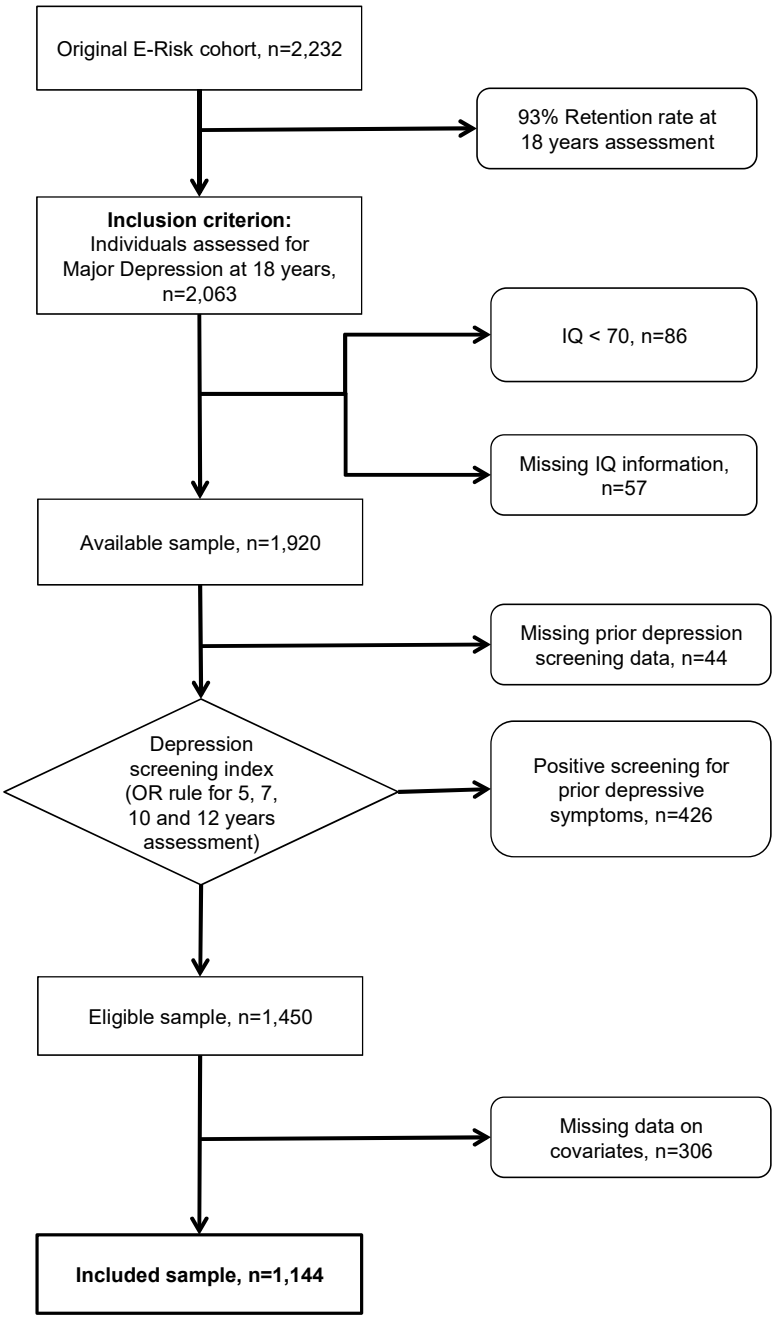
37. McGorry PD, Hartmann JA, Spooner R, Nelson B. Beyond the "at risk mental state" concept: transitioning to transdiagnostic psychiatry. *World Psychiatry.* 2018;17(2):133-142.
38. Fusar-Poli P, Solmi M, Brondino N, et al. Transdiagnostic psychiatry: a systematic review. *World Psychiatry.* 2019;18(2):192-207.
39. Caspi A, Moffitt TE. All for one and one for all: Mental disorders in one dimension. *American Journal of Psychiatry.* 2018;175(9):831-844.

Figure 1

Figure 1a-c.



b) E-Risk cohort



c) Dunedin cohort

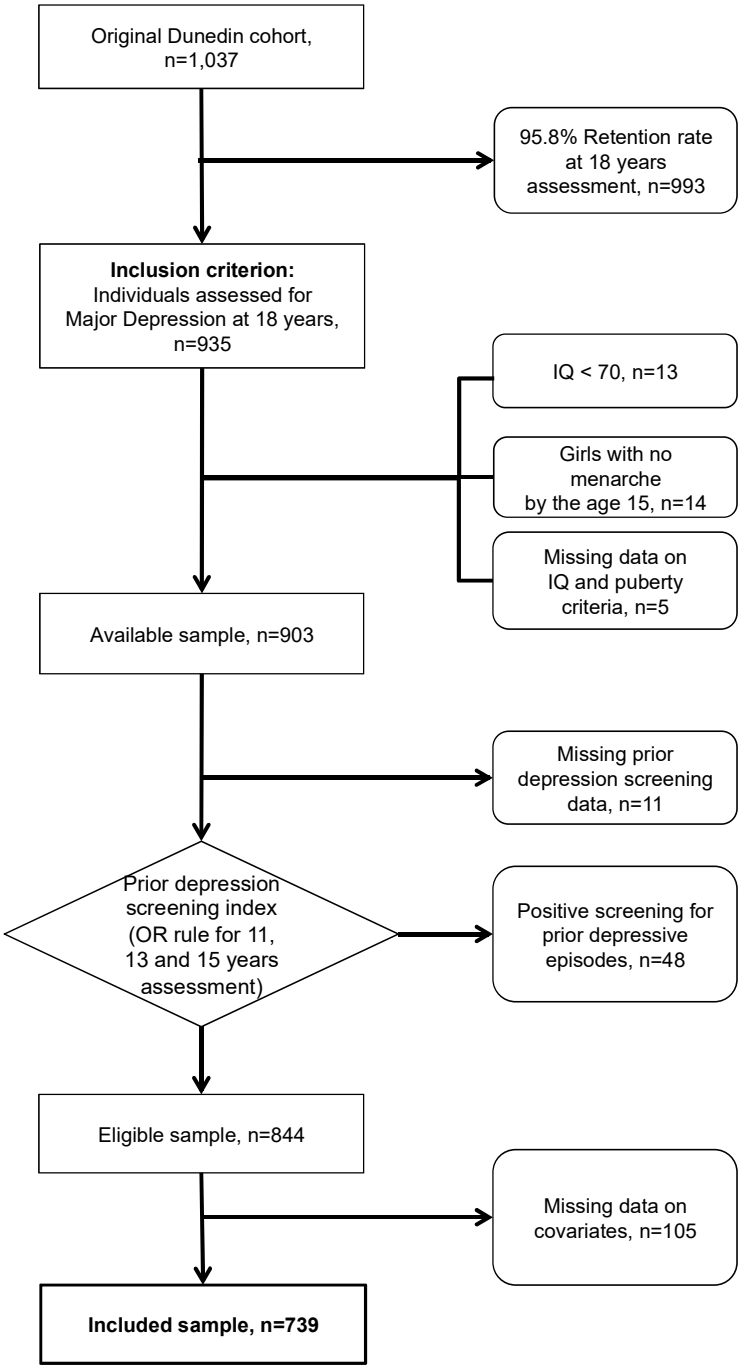


Figura 2

Figure 2.

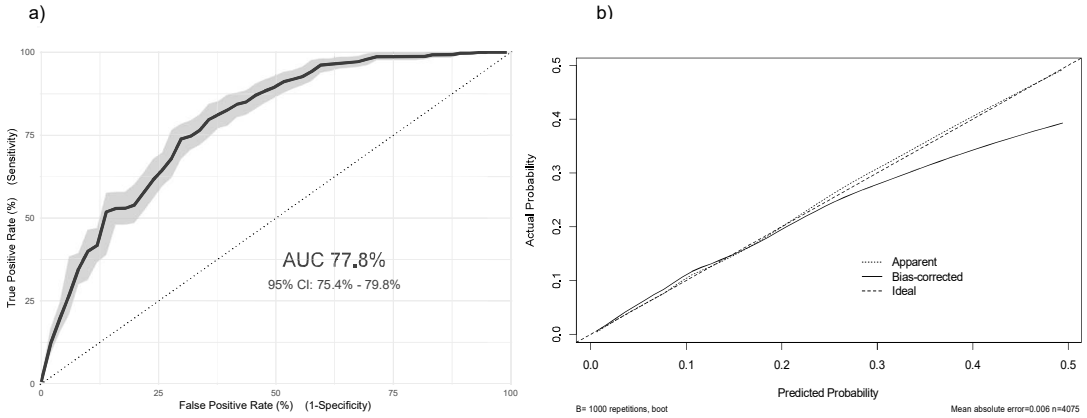
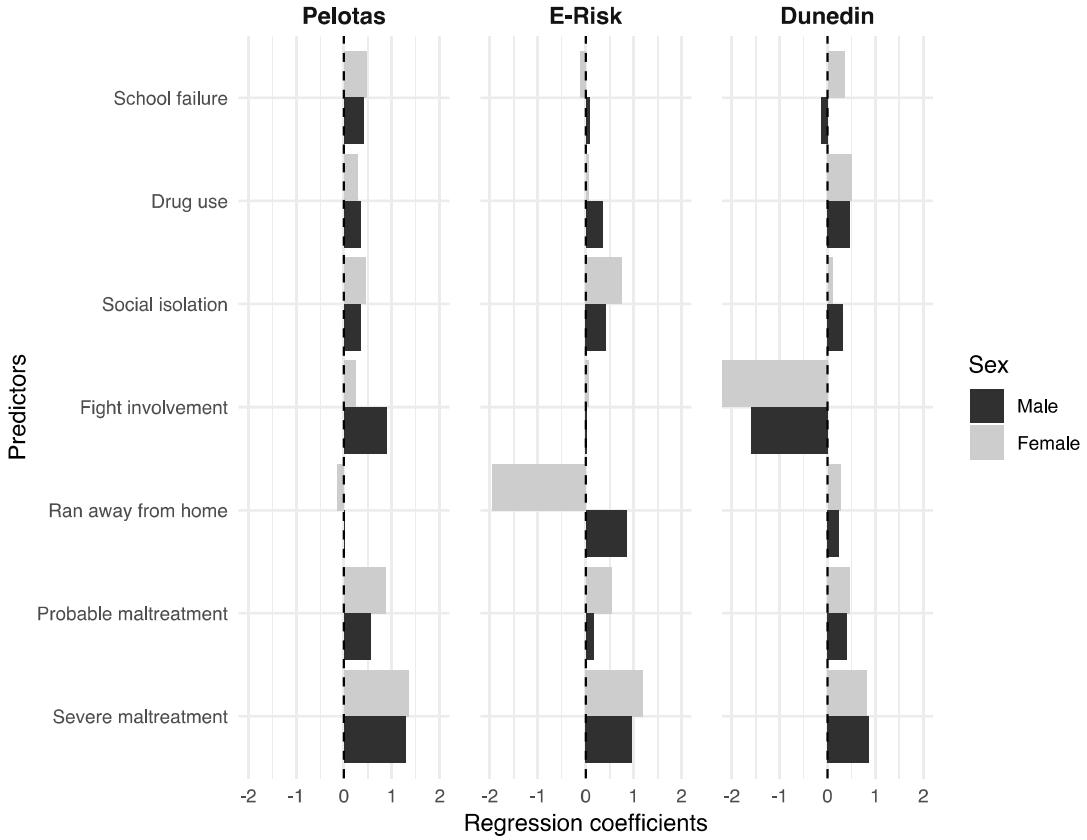


Figure 3.


Pelotas
E-Risk
Dunedin
School failure
Drug use
Social isolation
Fight involvement
Ran away from home
Probable maltreatment
Severe maltreatment
Predictors
-2
-1
0
1
2
Regression coefficients
Sex
Male
Female

ONLINE SUPPLEMENTARY MATERIAL

Supplement 1. Samples description

In 1993, all 5,249 children born live in the city of Pelotas were enrolled in the study. At the wave for ages 18-19 years old, the retention rate was 81.3% of the original sample. Further information on the cohort design can be found elsewhere.[1,2] Pelotas currently has around 300,000 inhabitants and is located at the extreme South of Brazil, near the Uruguayan border, a region with higher degrees of European genetic contributions when compared to other parts of Brazil. Its main economic activities are rice production, commerce and education. At the time of the beginning of the study, the infant mortality rate was 21 deaths per thousand births. Participants have been followed up at several points in time during infancy, childhood, adolescence and young adulthood. The original goals of the 1993 Cohort were to evaluate trends in maternal and child health indicators, through a comparison with results of the 1982 study, the first of a series of birth cohort studies performed in the city, to assess associations between early life variables and later outcomes, with particular emphasis on the detection of critical windows; and to improve data quality, using the lessons learned from the 1982 study.

The Environmental Risk (E-Risk) Longitudinal Twin study tracks the development of a nationally-representative birth cohort of 2,232 British twin children born in England and Wales in 1994-1995.[3] Families were recruited to represent the UK population of families with newborns in the 1990s, based on residential location throughout England and Wales and mothers' age. E-Risk families are representative of UK households across the spectrum of neighborhood-level deprivation.[4] The sample comprised 56% monozygotic and 44% dizygotic twin pairs, and sex was evenly distributed within zygosity (49% male). Follow-up home-visits were conducted when children were aged 7, 10, 12, and 18 years (participation rates were 98%, 96%, 96%, and 93%, respectively). The Joint South London and Maudsley and the Institute of Psychiatry Research Ethics Committee approved each phase of the study. Parents gave informed consent and twins gave assent between 5-12 years and then informed consent at age 18.

The Dunedin Study represents the full range of socioeconomic status on NZ's South Island and matches the NZ National Health and Nutrition Survey on key health indicators (e.g., BMI, smoking, GP visits).[5] The cohort is primarily white; fewer than 7% self-identify as having non-Caucasian ancestry, matching the South Island. Assessments were carried out at birth and ages 3, 5, 7, 9, 11, 13, 15, 18, 21, 26, 32, and, most recently, 38 years, when 95% of the 1,007 study members still alive took part. At each assessment, each study member is brought to the research unit for a full day of interviews and examinations. These data are supplemented by questionnaires completed by persons who know the study members well and by official record searches.

Table S1. Comparative table of variables' definitions and assessment strategies among the included studies

Variable (reference category)[a]	Pelotas	E-Risk	Dunedin
Previous depression screening	Any evidence ("OR" rule) of depressive symptoms, assessed by borderline score in emotional SDQ subscale for self and parent assessment at age 11 and for parent assessment at age 15	Any evidence ("OR" rule) of depressive symptoms, assessed at ages 5, 7 and 10 by a depression subscale derived from a combination of mother and teacher CBCL for emotional problems, using the 93th percentile as cutpoint, and at age 12 by self-reported CDI scores (with a clinical cut-off >= 20)	Any positive depressive episode assessment using the DISC at ages 11, 13 or 15
Puberty	Second stage of Tanner classification at age 15	Not applied in the sample (see main text)	Menarche for girls at age 15[b]
Sex (Male)	Self reported sex	Self reported sex	Self reported sex
Skin color (White)	Self assigned skin color	Self assigned skin color	Self assigned ancestry
Childhood maltreatment (None)	Responses to seven dichotomous questions regarding lifetime psychological, physical and sexual abuse and/or neglect at age 15;[6] zero positive answers=none, 1 positive=probable, 2 or more=severe; inserted as a categorical variable into the model	Prospectively obtained variable for sexual/physical abuse up to age 12 based on mother reports, researcher observations, and social services referral information, coded as none, probable, or definite; inserted as a categorical variable into the model	Prospectively obtained variable for childhood maltreatment up to age 12 based on mother reports, researcher observations, social services referral information, and retrospective self-reports, coded none/probable/severe; inserted as a categorical variable into the model
School failure (No)	In Brazil, if a child fails to achieve a predetermined score at the end of the school year, the child is retained by the school to repeat the same school year; positive answers to the dichotomous question: "Have you ever been retained in school?" was classified as "failing at school"=1; otherwise=0	Evaluation of sample's distribution of English/Math performance at age 12, considering those below the 20th percentile as "failing at school"=1; otherwise=0	Those who left school at age 15 with no qualifications were classified as "failing at school"=1; otherwise=0
Social isolation (0)	Responses to the question: "Do you normally meet up with friends to chat, play or do other things? If YES, how many days in a given week?" were dichotomized as follows: responses "no" were classified as "1" and "yes" were classified as "0"	Combination of CBCL and TRF items on social isolation was pooled into a 3 strata categorical social isolation variable (low, moderate and high social isolation)[7], and then reclassified into a dichotomized variable: high social isolation=1 and low/moderate=0	Those below the 10th percentile of the sample's distribution in the Peer Attachment Scale were classified as "socially isolated"=1; those above it were classified as "socially isolated"=0
Fights (No)	Responses to the dichotomous question: "In the last year, have you ever gotten into a physical fight that someone got hurt?"; positive answers were classified as "1", and negative as "0"	Combination ("AND" rule) of two dichotomous questions: "Do you sometimes hit someone when you are having an argument?" and "Do you sometimes start fights with people?"; positive answers to both questions were classified as "1"; otherwise, "0"	Response to the question: "In the past year, how many times have you fought in the street or other public place?"; regrouped into a dichotomized variable: once or more="1" and never="0"
Ran away (No)	Responses to the dichotomous question: "Have you ever run away from home?"; positive answers were classified as "1", and negative as "0"	Responses to the dichotomous question: "Have you run away from home and stayed away for the night?"; positive answers were classified as "1", and negative as "0"	Question about running away overnight; regrouped into a dichotomized variable: once or more="1" and never="0"

Drug use (No)	Dichotomous variable combining responses to dichotomous questions about any lifetime use of alcohol, tobacco, cannabis, cocaine and inhalants; any positive answer="1"; otherwise="0"	Dichotomous variable combining responses to dichotomous questions about any lifetime use of alcohol, tobacco, cannabis, pills and inhalants; any positive answer="1"; otherwise="0"	Dichotomous variable combining responses to dichotomous questions about any lifetime use of alcohol, tobacco, cannabis, inhalants and other illegal drug; any positive answer="1"; otherwise="0"
Relationship with mother (Great)	Question: "How do you rate your relationship with your mother?"; choices of answers: great=1, very good=2, good=3, regular=4, bad=5; inserted as a categorical variable into the model	No match	Continuous variable obtained from responses to the Parent Attachment Scale[c]
Relationship with father (Great)	Question: "How do you rate your relationship with your father?"; choices of answers: great=1, very good=2, good=3, regular=4, bad=5; inserted as a categorical variable into the model	No match	
Relationship between parents (Great)	Question: "How do you rate the relationship between your father and your mother?"; choices of answers: great=1, very good=2, good=3, regular=4, bad=5; inserted as a categorical variable into the model	No match	Parents arguing frequency; choice of answers: never=1, one or two times=2, sometimes=3, often=4, all the time=5; inserted into the model as a categorical variable
Depression diagnosis at age 18 (No)	Evaluation of major depressive episode diagnosis with DSM-IV-TR criteria in the previous two weeks[d]	Evaluation of major depressive episode diagnosis with DSM-IV criteria in the previous 12 months	Past-year major depressive episode using DIS at age 18 following the DSM-III-R criteria

[a]For variables included as predictors.

[b]No puberty assessment available for boys in the cohort.

[c]See Methods S2 for details.

[d] All psychiatric diagnoses were assessed by trained psychologists using an instrument derived from the Mini International Neuropsychiatric Interview.[8]

CBCL: Child Behavior Checklist; CDI: Children's Depression Inventory; DIS: Diagnostic Interview Schedule; DISC: Diagnostic Interview Schedule for Children; DSM: Diagnostic and Statistical Manual of Mental Disorders; SDQ: Strengths and Difficulties Questionnaires; TRF: Teacher's Report Form.

For comparison purposes, the harmonization among the cohorts of the selected variables was performed a priori, when consensus was reached with each cohort's primary investigator before datasets sharing and assessment.

Table S2. Coefficients comparison for the selected Pelotas model in complete-case analyses and after multiple imputation strategy

	Complete-case analyses	Multiple imputation
Intercept	-4.642	-4.687
Sex	0.325	0.270
Skin color	-0.030	-0.096
School failure	0.290	0.442
Drug use	0.121	0.156
Social isolation	0.127	0.147
Fights involvement	0.580	0.588
Ran away from home	-0.017	0.028
Probable maltreatment	0.422	0.362
Severe maltreatment	0.652	0.755
Relat. w/mother=2	0.054	0.133
mother=3	0.161	0.209
mother=4	-0.026	0.009
mother=5	0.006	0.024
Relat. w/father=2	-0.010	0.005
father=3	0.297	0.359
father=4	0.181	0.205
father=5	0.237	0.320
Relat. bw parents=2	-0.004	-0.017
parents=3	0.251	0.186
parents=4	0.163	0.194
parents=5	0.037	0.130
Sex*Skin color	0.170	0.110
Sex*School failure	0.114	0.097
Sex*Drug use	0.108	0.130
Sex*Social isolation	0.269	0.222
Sex*Fights	-0.395	-0.376
Sex*Ran away	-0.101	-0.211
Sex*Probable maltreatment	0.369	0.411
Sex*Severe maltreatment	0.382	0.313
Sex*Relat. w/mother=2	-0.219	-0.206
Sex*mother=3	0.060	0.029
Sex*mother=4	-0.203	-0.239
Sex*mother=5	-1.293	-1.386
Sex*Relat. w/father=2	0.143	0.154
Sex*father=3	-0.279	-0.201
Sex*father=4	-0.103	-0.150
Sex*father=5	0.537	0.416
Sex*Relat. bw parents=2	0.011	0.079
Sex*parents=3	-0.035	-0.094
Sex*parents=4	0.158	0.118
Sex*parents=5	0.075	0.149

Complete-case analyses: Coefficients obtained from a penalized logistic regression model using penalized maximum likelihood estimation (PMLE) after list-wise exclusion due to missing data on covariates. Multiple imputation: Coefficients obtained after multiple imputation using 10 imputation datasets and 10 iterations using the method of chained equations (R package *mice*). Frequency of missing data for each parameter (from highest to lowest): Relationship with father: 93; Relationship between parents: 81; Childhood maltreatment: 59; Ran away: 46; Relationship with mother: 43; Drug use: 29; Fights: 19. Relat. w/mother: adolescent's relationship with his(her) mother; Relat. w/father: adolescent's relationship with his(her) father; Relat. bw parents: relationship between parents (for all three, reference category 1="great"). Interaction term is shown as "variable"*"variable".

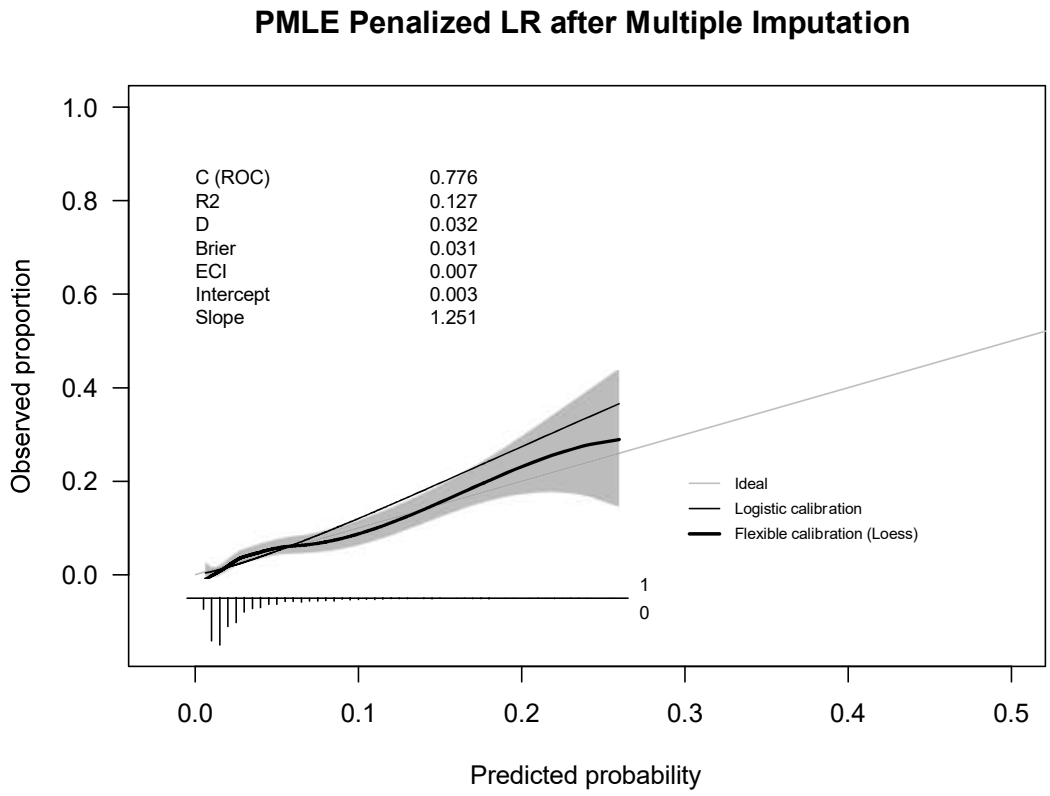


Figure S1. Predictive performance of the Pelotas model after multiple imputation

Calibration indicates the agreement between predicted probabilities and observed frequencies. A calibration plot is a graphical assessment of calibration, where the calculated predicted probability is shown on the x-axis, and observed frequency of the outcome on the y-axis. The flexible calibration line shows the result of a smoothing technique (Loess algorithm) used to estimate the observed probabilities of the binary outcome in relation to the predicted probabilities.

The grey shaded area indicates the 95% confidence interval around the estimated calibration line. The logistic calibration curve shows the result of the estimation of the observed proportions when a logistic model is used for the outcome as a function of the linear predictor of the developed model. The 45-degree line (labelled 'Ideal') has slope=1 and indicates perfect calibration. PMLE: Penalized maximum likelihood estimation; LR: Logistic regression. C (ROC): C-statistic; R2: Nagelkerke's R^2; D: Discrimination index; Brier: Brier score; ECI: Estimated Calibration Index; Intercept: Regression intercept of linear predictor; Slope: Regression slope of linear predictor.

Table S3. Pelotas sample characteristics for individuals included and excluded from the final analyses

	Included	Excluded	Total[a]
Male sex	977 (44.6%)[b]	1006 (54.1%)[c]	1983 (48.9%)
White skin color	1478 (67.4%)[b]	1024 (59.9%)[c]	2502 (64.1%)
Childhood maltreatment			
No	1539 (70.2%)[b]	914 (62.5%)[c]	2453 (67.1%)
Probable	390 (17.8%)[b]	274 (18.7%)[b]	664 (18.2%)
Severe	263 (12.0%)[b]	275 (18.8%)[c]	538 (14.7%)
School failure	1127 (51.4%)[b]	1301 (76.1%)[c]	2428 (62.2%)
Social isolation	231 (10.5%)[b]	266 (15.5%)[c]	497 (12.7%)
Any drug use	1367 (62.4%)[b]	992 (61.5%)[b]	2359 (62.0%)
Fights	211 (9.6%)[b]	242 (14.9%)[c]	453 (11.9%)
Ran away	80 (3.7%)[b]	108 (6.8%)[c]	188 (5.0%)
Relationship with mother			
Great	1417 (64.6%)[b]	853 (53.4%)[c]	2270 (59.9%)
Very good	430 (19.6%)[b]	342 (21.4%)[b]	772 (20.4%)
Good	264 (12.0%)[b]	257 (16.1%)[c]	521 (13.8%)
Regular	68 (3.1%)[b]	119 (7.5%)[c]	187 (4.9%)
Bad	13 (0.6%)[b]	25 (1.6%)[c]	38 (1.0%)
Relationship with father			
Great	1019 (46.5%)[b]	606 (40.3%)[c]	1625 (44.0%)
Very good	434 (19.8%)[b]	269 (17.9%)[b]	703 (19.0%)
Good	370 (16.9%)[b]	290 (19.3%)[b]	660 (17.9%)
Regular	237 (10.8%)[b]	218 (14.5%)[c]	455 (12.3%)
Bad	132 (6.0%)[b]	120 (8.0%)[c]	252 (6.8%)
Relationship between parents			
Great	886 (40.4)%[b]	520 (34.3%)[c]	1406 (37.9%)
Very good	421 (19.2%)[b]	288 (19.0%)[b]	709 (19.1%)
Good	404 (18.4%)[b]	285 (18.8%)[b]	689 (18.6%)
Regular	301 (13.7%)[b]	243 (16.0%)[c]	544 (14.7%)
Bad	180 (8.2%)[b]	180 (11.9%)[c]	360 (9.7%)
Depressive episode	69 (3.1%)[b]	93 (5.0%)[c]	162 (4.0%)
Total	2,192	1,860	4,052

[a] Total number of individuals meeting our inclusion criteria. The sum of included and excluded samples differs from the total number shown in some rows due to 457 individuals with missing data for exclusionary criteria. Categorical variables presented as percentages (according to column). Results derived from a chi-square (χ^2) test. Superscript letters "b" and "c" denote column differences between included and excluded samples: different letters show significant and equal letters indicate non-significant differences from each other at a 0.05 level.

Table S4. Variables' regression coefficients for each developed model from the Pelotas dataset

	LR	PMLE	Ridge	0.25	0.50	0.75	LASSO
Intercept	-5.879	-4.642	-4.313	-3.963	-3.948	-3.978	-3.961
Sex	1.569	0.325	0.137	0.047	0.002	.	.
Skin color	-0.284	-0.030	0.004	.	.	.	.
School failure	0.863	0.290	0.152	0.034	.	.	.
Drug use	0.160	0.121	0.081	.	.	.	.
Social isolation	0.673	0.127	0.125	.	.	.	.
Fights involvement	1.582	0.580	0.491	0.336	0.350	0.364	0.330
Ran away from home	-0.539	-0.017	0.001	.	.	.	.
Probable maltreatment	0.148	0.422	0.210	0.063	.	.	.
Severe maltreatment	1.107	0.652	0.509	0.612	0.767	0.890	0.955
Relat. w/mother=2	0.894	0.054	0.025	.	.	.	.
mother=3	0.045	0.161	0.197	0.232	0.234	0.233	0.216
mother=4	-6.321	-0.026	-0.185	.	.	.	.
mother=5	-6.430	0.006	-0.596	.	.	.	.
Relat. w/father=2	-1.103	-0.010	-0.091	.	.	.	.
father=3	1.242	0.297	0.177	.	.	.	.
father=4	0.811	0.181	0.107	.	.	.	.
father=5	1.258	0.237	0.302	0.107	.	.	.
Relat. bw parents=2	-0.356	-0.004	-0.065	.	.	.	.
parents=3	0.365	0.251	0.177	0.007	.	.	.
parents=4	-0.010	0.163	0.127	.	.	.	.
parents=5	-0.897	0.037	-0.036	.	.	.	.
Sex*Skin color	0.488	0.170	0.144	0.045	0.021	0.012	.
Sex*School failure	-0.463	0.114	0.169	0.201	0.257	0.279	0.283
Sex*Drug use	0.093	0.108	0.124	0.075	0.058	0.051	0.038
Sex*Social isolation	-0.189	0.269	0.236	0.169	0.166	0.177	0.147
Sex*Fights	-1.615	-0.395	-0.250	.	.	.	.
Sex*Ran away	0.386	-0.101	-0.061	.	.	.	.
Sex*Probable maltreatment	0.966	0.369	0.399	0.505	0.651	0.710	0.714
Sex*Severe maltreatment	0.240	0.382	0.379	0.287	0.188	0.092	0.025
Sex*Relat. w/mother=2	-1.337	-0.219	-0.126	.	.	.	.
Sex*mother=3	0.068	0.060	0.076	.	.	.	.
Sex*mother=4	5.839	-0.203	-0.013	.	.	.	.
Sex*mother=5	-1.361	-1.293	-0.716	-0.110	-0.039	-0.056	.
Sex*Relat. w/father=2	1.260	0.143	0.150	.	.	.	.
Sex*father=3	-1.499	-0.279	-0.164	.	.	.	.
Sex*father=4	-0.949	-0.103	-0.013	.	.	.	.
Sex*father=5	-0.492	0.537	0.459	0.565	0.691	0.713	0.715
Sex*Relat. bw parents=2	0.304	0.011	0.039	.	.	.	.
Sex*parents=3	-0.162	-0.035	0.009	.	.	.	.
Sex*parents=4	0.347	0.158	0.161	0.111	0.089	0.084	0.066
Sex*parents=5	0.890	0.075	0.129	.	.	.	.

LR: Logistic regression model; PMLE: penalized maximum likelihood estimation logistic regression; 0.25: Elastic-Net with alpha=0.25; 0.50: Elastic-Net with alpha=0.50; 0.75: Elastic-Net with

alpha=0.75; Relat. w/mother: adolescent's relationship with his(her) mother; Relat. w/father: adolescent's relationship with his(her) father; Relat. bw parents: relationship between parents (for all three, reference category 1="Great"). Interaction term is shown as "variable"*"variable".

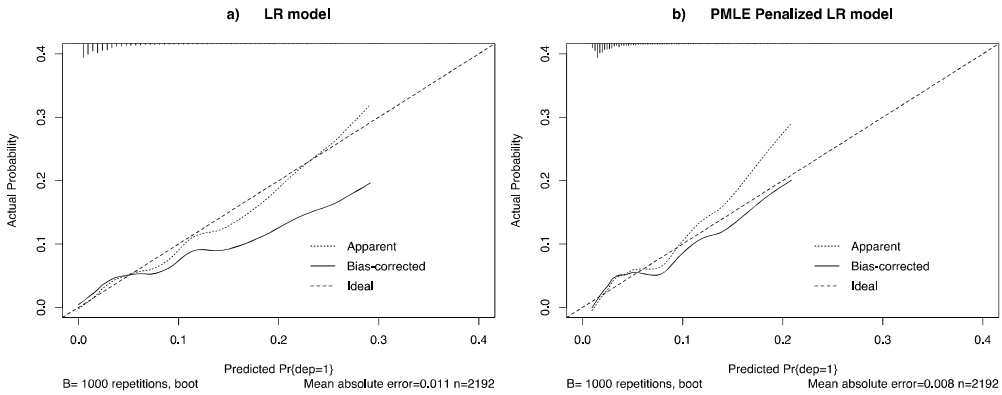


Figure S2a-b. Calibration plots after internal validation using 1,000 iterations bootstrapping for: a) the logistic regression (LR), and b) the Penalized maximum likelihood estimation (PMLE) LR model.

Apparent and bias-corrected results were plotted as a nonparametric calibration curve, estimated over a sequence of predicted values vs. observed values using a smoothing technique.

Supplement 2. Complementary description of statistical analysis

From a predictive modeling perspective, the use of the full-model, the one that contains all pre-specified terms, will usually be the one that predicts most accurately in new data (model generalization or validation).[9] However, the maintenance of all pre-selected variables in the final model can sometimes produce overfitting. Overfitting can be described as an overly optimistic predictive estimate of accuracy, biasing the developed model's generalization. Several statistical strategies have been developed to deal with overfitting, where a penalization factor is included into the model for shrinkage of coefficients and/or variable selection.[9-11] Shrinkage (or Penalization) strategies such as Penalized Maximum Likelihood Estimation (PMLE), Ridge regression, LASSO (Least Absolute Shrinkage and Selection Operator), and Elastic-Net penalization have been used in the literature for reducing overfitting and variance, optimizing prediction on new data.[9] The Elastic-Net machine learning technique blends both Ridge and LASSO penalization approaches,[12] having the advantage of combining shrinkage of parameters with selection of meaningful variables, excluding non-contributing factors without the risks of p value-based variable selection techniques.[9,13]

Following current recommendations of transparent reporting,[13,14] we describe here a stepwise description of the strategy for statistical analysis:

Step one:

For the Pelotas sample, no transformation or data handling was performed, keeping the variables as they were originally collected. All predictors were entered as categorical variables into the models. Given that biological sex is a consolidated risk factor for depression,[15] we have included into the models interaction terms of sex with all other selected variables, as well as their main effects. The penalty factor used in the PMLE was empirically obtained from our data. For the Elastic-Net approach, we have *a priori* defined a grid of values for the hyperparameter alpha, ranging from 0 (full Ridge) to 1 (full LASSO), with increments of 0.25. For each alpha value, a 10-fold cross-validation was used to select the penalty coefficient (lambda) that minimized the mean squared prediction error, which was then used for shrinkage of coefficients and/or variable selection. The Elastic-Net models with alphas > 0 had some variables excluded from final analyses, as their contributions after penalization were negligible (see Table S4).

Step two:

Discrimination and calibration metrics, paired with calibration plots, are the commonest performance measures for model assessment.[13,16] The discriminative capacity of the model is defined as the ability to differentiate low- from high-risk individuals. For binary outcomes, discrimination can be quantified by the AUC, also named concordance statistic (C-statistic). Calibration refers to the agreement of predicted and observed outcomes, and can be investigated by calibration plots using flexible calibration curves or categorizations of predicted risk. Additionally to models' discrimination and calibration measures, results of explained variation (R^2), model's goodness-of-fit (Likelihood Ratio χ^2), predictive accuracy (Brier score), the discrimination index (D), and the maximum error in predicted probabilities (Emax) were also calculated.[9,16]

Step three:

A predictive model generally performs better in the sample used to develop the model, namely apparent performance, compared to other samples. This performance measures' discrepancy is called model optimism. Assessment of optimism is crucial for estimating the chance of a model's reproducibility and implementation. We have submitted the models to 1,000 resampling iterations bootstrapping, the most recommended internal validation process,[9,17] to get bias(overfitting)-corrected estimates. Then, this bias-corrected model was plotted as a nonparametric calibration curve, estimated over a sequence of predicted values vs. observed values using a smoothing technique.

Considering the non-matching variables in the E-Risk and Dunedin samples, there was a need to rebuild the Pelotas model for each of the external validation samples, recalculating the linear predictor considering only the variables available for each comparison. Also, strategies for enhancing comparability between datasets were implemented. Since there was only a combined parent relationship assessment variable in the Dunedin sample, with no adequate matching to the separated

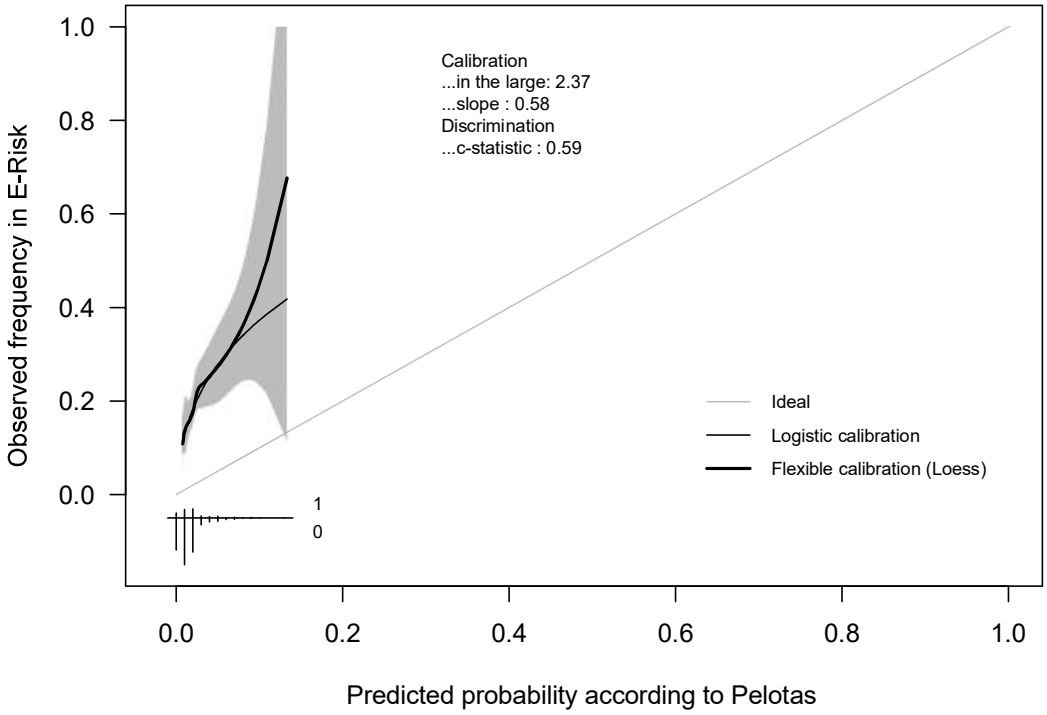
evaluation of adolescent relationship with each parent, we have decided to transform the Pelotas variables for better comparison to the Dunedin dataset. After transforming both variables (relationship with mother and relationship with father) into continuous variables, we have derived the arithmetical mean of both variables (summed them and divided by two). As the Dunedin study used the Parent Attachment Scale, a different evaluation strategy, with a larger range of results, with higher results suggesting a better relationship, we have also Z-scored both the obtained arithmetical mean and the Parent Attachment Scale results for better comparability. Given that the Pelotas coding of parent relationship assessment variables have an opposite direction (higher value, worse relationship) to the Dunedin variable, we have multiplied its coefficient value by -1. As mentioned in the main text, there were no family relationship assessment available in the E-Risk dataset for comparison.

Model update:

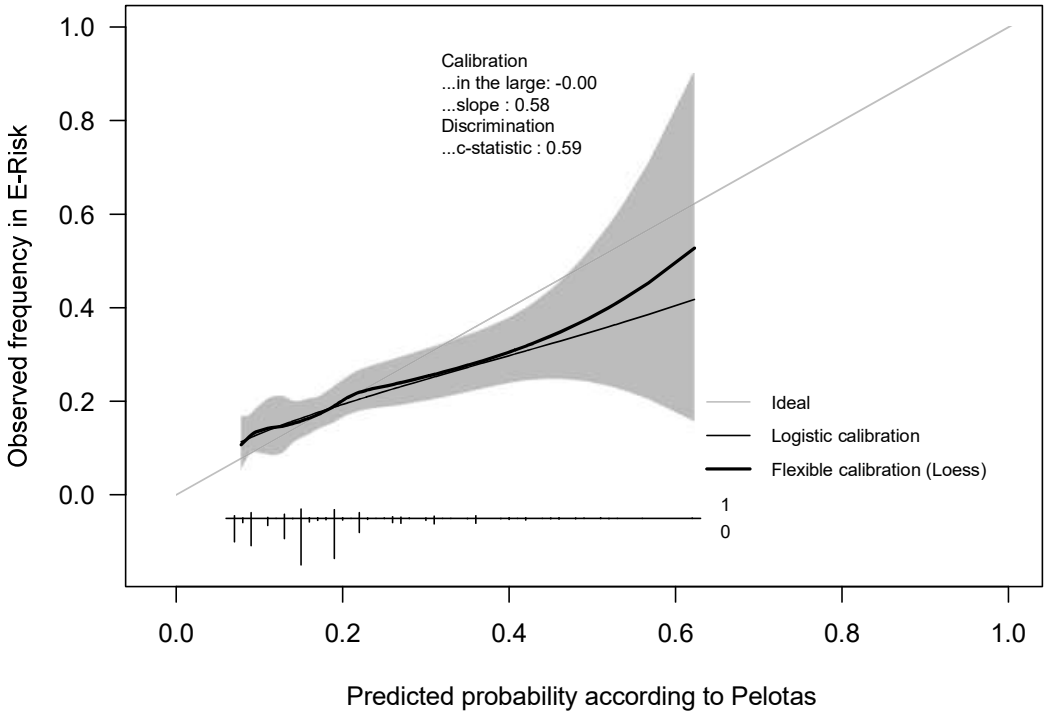
Current methodological guidelines recommend the identification of calibration-in-the-large problems should have priority when evaluating external validation, since miscalibration can cause systematically wrong decision making. Model updating (or adjustment) intends to make the average predicted probability equal to the observed overall event rate by fitting a new logistic regression model in the validation sample using the new intercept as the only free parameter, with the original linear predictor (obtained from the development sample) as an offset variable.

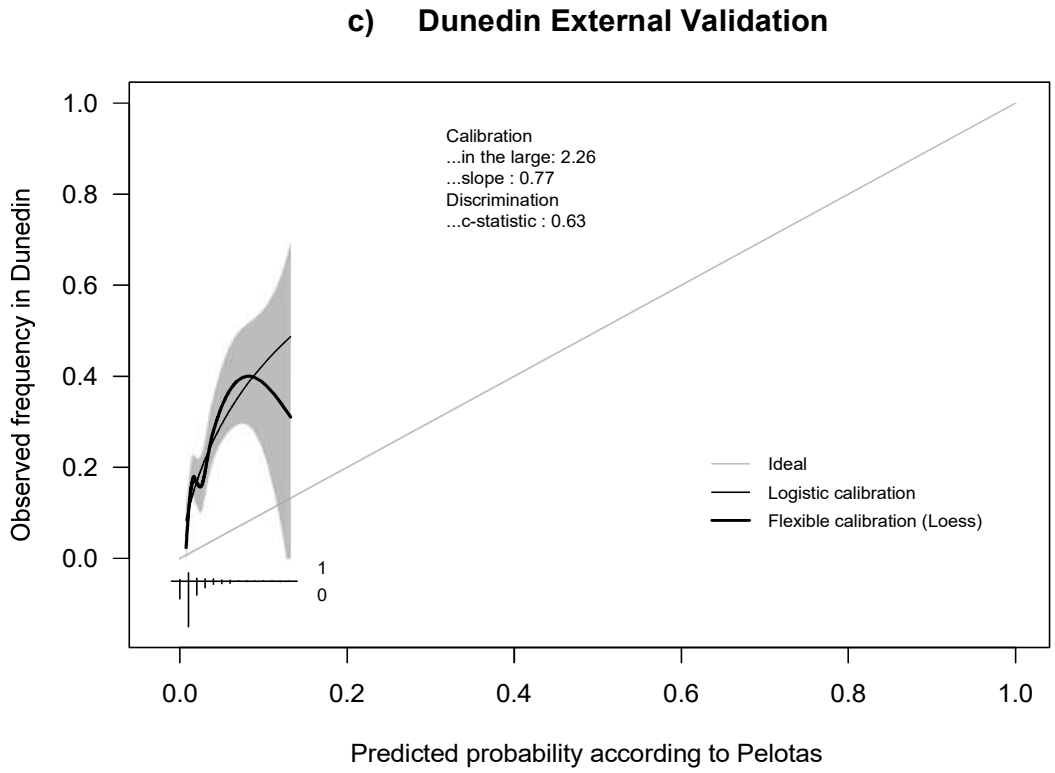
The adjustment of Pelotas model by correcting its intercept for each cohort, resulted in an improvement of calibration measures. In the E-Risk dataset, this adjustment reduced the Brier score by 14.3% (from 0.17 to 0.14) and the Emax value by 11.2% (from 0.29 to 0.26), keeping the remaining metrics unchanged. A similar reduction was also achieved in Dunedin: 14.6% and 66.6% for Brier and Emax scores, respectively.

a) E-Risk External Validation



b) Adjusted E-Risk External Validation





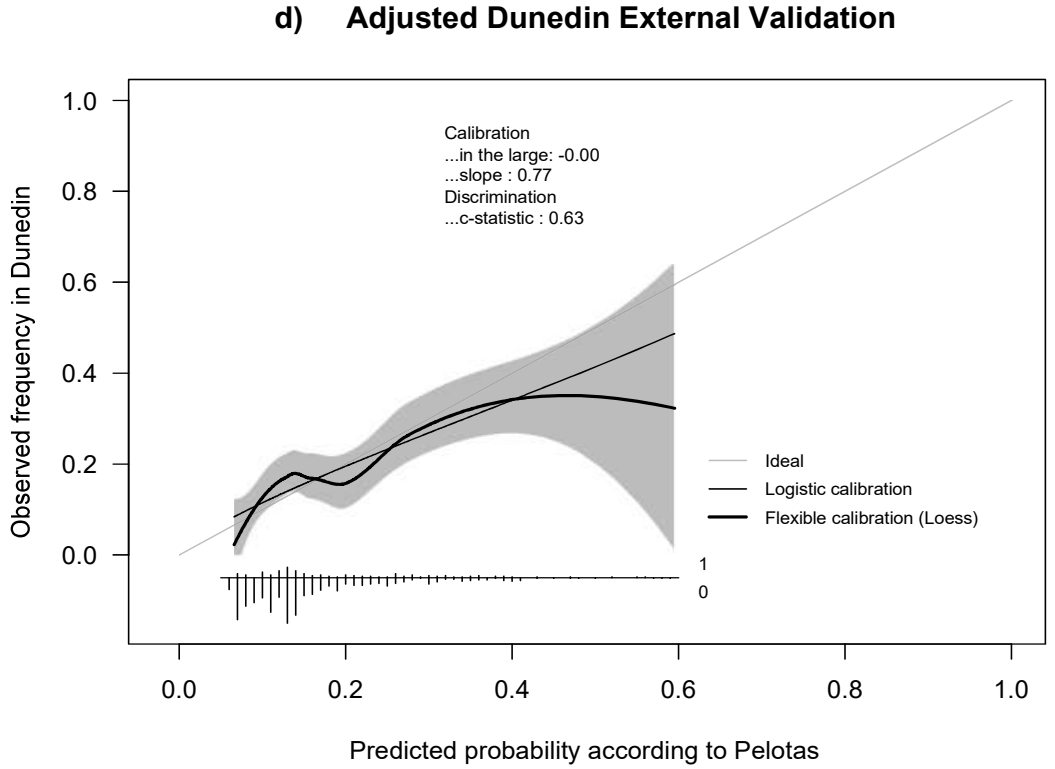


Figure S3a-d. Calibration plots for standard and adjusted (updated) external validation assessment in each sample.

Calibration indicates the agreement between predicted probabilities and observed frequencies. A calibration plot is a graphical assessment of calibration, where the calculated predicted probability is

shown on the x-axis, and observed frequency of the outcome on the y-axis. The flexible calibration line shows the result of a smoothing technique (Loess algorithm) used to estimate the observed probabilities of the binary outcome in relation to the predicted probabilities.
The grey shaded area indicates the 95% confidence interval around the estimated calibration line. The logistic calibration curve shows the result of the estimation of the observed proportions when a logistic model is used for the outcome as a function of the linear predictor of the developed model. The 45-degree line (labelled 'Ideal') has slope=1 and indicates perfect calibration. The distribution of predicted probabilities is shown at the bottom of the graphs.
Calibration in the large: compares the mean of all predicted risks with the mean observed risk. Calibration slope: measure of agreement between observed and predicted risk of the outcome across the whole range of predicted values. C-statistic: Concordance statistic, equal to the area under the curve of the receiver operating characteristic (AUC-ROC) in binary endpoints.

Supplement 3. Model validation analyses

For evaluation of relatedness of the development and validation samples, we have followed current recommendations for assessment of disparities in case mix.[18,19] Firstly, we compared the distribution of predictors of the validated model, as well as the outcome of interest, among samples (Table 1). Secondly, we have evaluated the dispersion of the model's predicted risk in the development and validation samples. Assessing the spread and mean value of the linear predictor (lp) of the derived model in the development and validation samples, in the Pelotas sample, the mean value of lp was -3.678, with a spread, measured by the standard deviation (SD), of 0.685. In the E-Risk sample, the mean lp was -4.012 and SD was 0.562. For the Dunedin sample, the mean lp was -3.983, with a SD of 0.601. Finally, we assessed predictors' coefficients for the variables available in all cohorts after merging all datasets into an aggregated sample, rebuilding the PMLE model after including each cohort's main-effects and their interaction terms with all predictors (cohort*predictor), as can be seen in Table S5 and Figure 3.

Additionally, following recommended statistical strategies,[20] we quantified the impact of differences in case-mix on the model's validation performance. For this analysis, we assume that the regression coefficients for assessed predictors and the model intercept are fully correct for the validation setting. Simulating the outcome from the observed case-mix in the development sample, assuming the prediction model is correct for the new sample, differences in performance between the development and validation assessments suggest real differences in the regression coefficients' weights.

We have also calculated the performance obtained by refitting the model in the validation data, estimating coefficients that would be optimal for the validation data. This approach provides an upper bound for model performance if the coefficients from the development setting were exactly equal to those in the validation setting. As this upper bound is a result of both case-mix and the effects of predictors in the validation setting, differences in performance results could be related to both.[20]

Table S5. Exploratory analysis after merging all datasets from Pelotas, E-Risk and Dunedin cohorts for variables available in all cohorts[a]

	Coefficient	SE
Intercept	-5.027	0.355
Sex	0.857	0.301
School failure	0.407	0.295
Drug use	0.332	0.261
Social isolation	0.341	0.470
Fights involvement	0.882	0.394
Ran away from home	0.005	0.653
Probable Child. Mt	0.544	0.340
Severe Child. Mt	1.293	0.403
E-Risk cohort	2.749	0.354
Dunedin cohort	2.345	0.404
Sex*School failure	0.053	0.273
Sex*Drug use	-0.047	0.230
Sex*Social isolation	0.117	0.460
Sex*Fights	-0.640	0.514
Sex*Ran away	-0.131	0.662
Sex*Probable Child. Mt	0.334	0.330
Sex*Severe Child. Mt	0.048	0.403
Sex*E-Risk	-0.026	0.289
Sex*Dunedin	-0.094	0.355
School failure*E-Risk	-0.329	0.351

School failure*Dunedin	-0.535	0.496
Drug use*E-Risk	0.018	0.274
Drug use*Dunedin	0.133	0.298
Social isolation*E-Risk	0.086	0.567
Social isolation*Dunedin	-0.028	0.566
Fights*E-Risk	-0.854	0.445
Fights*Dunedin	-2.483	1.490
Ran away*E-Risk	0.844	1.026
Ran away*Dunedin	0.227	0.750
Probable Child. Mt*E-Risk	-0.387	0.419
Severe Child. Mt*E-Risk	-0.330	0.558
Probable Child. Mt*Dunedin	-0.130	0.396
Severe Child. Mt*Dunedin	-0.419	0.559
Sex*School failure*E-risk	-0.250	0.412
Sex*School failure*Dunedin	0.421	0.623
Sex*Drug use*E-Risk	-0.245	0.301
Sex*Drug use*Dunedin	0.085	0.345
Sex*Social isolation*E-Risk	0.219	0.674
Sex*Social isolation*Dunedin	-0.328	0.761
Sex*Fights*E-Risk	0.675	0.665
Sex*Fights*Dunedin	-1.069	3.249
Sex*Ran away*E-Risk	-2.660	3.462
Sex*Ran away*Dunedin	0.169	0.835
Sex*Probable Child. Mt*E-Risk	0.049	0.479
Sex*Severe Child. Mt*E-Risk	0.193	0.746
Sex*Probable Child. Mt*Dunedin	-0.281	0.459
Sex*Severe Child. Mt*Dunedin	-0.099	0.653

[a] Due to the requirement of availability of variables in all datasets for this exploratory analysis, Pelotas model's coefficients had their values recalculated, as some previously included variables were excluded for comparability.

Interaction term is shown as "variable"*"variable" or "variable"*"variable"*"variable".

Mt: maltreatment. SE: standard error.

To calculate the coefficients shown in Figure 3 of the manuscript, the above coefficient values were summed considering variables sex and cohort, having male sex and Pelotas cohort as references. For instance, for estimation of the coefficient corresponding to the "Fights involvement" variable for females in the Dunedin cohort, this variable's main-effect (0.882) was summed to the interaction term sex*Fights (-0.640), to the interaction term Fights*Dunedin (-2.483), and to the interaction term sex*fights*Dunedin (-1.069), resulting a coefficient value of -3.309.

Table S6. Sensitivity analysis assessing the impact of the exclusion criteria on the Pelotas model

	Included sample (n=2,192)	**Available sample** (n=3,290)
Intercept	-4.642	-4.564
Sex	0.325	0.416
Skin color	-0.030	-0.107
School failure	0.290	0.205
Drug use	0.121	0.069
Social isolation	0.127	0.031
Fights involvement	0.580	0.480
Ran away from home	-0.017	-0.119
Probable maltreatment	0.422	0.420
Severe maltreatment	0.652	0.582
Relat. w/mother=2	0.054	-0.006
mother=3	0.161	0.188
mother=4	-0.026	0.029
mother=5	0.006	0.041
Relat. w/father=2	-0.010	0.040
father=3	0.297	0.332
father=4	0.181	0.122
father=5	0.237	0.172
Relat. bw parents=2	-0.004	0.001
parents=3	0.251	0.241
parents=4	0.163	0.245
parents=5	0.037	0.014
Sex*Skin color	0.170	0.169
Sex*School failure	0.114	0.098
Sex*Drug use	0.108	0.025
Sex*Social isolation	0.269	0.463
Sex*Fights	-0.395	-0.649
Sex*Ran away	-0.101	-0.168
Sex*Probable maltreatment	0.369	0.332
Sex*Severe maltreatment	0.382	0.504
Sex*Relat. w/mother=2	-0.219	-0.246
Sex*mother=3	0.060	0.210
Sex*mother=4	-0.203	-0.023
Sex*mother=5	-1.293	0.598
Sex*Relat. w/father=2	0.143	0.167
Sex*father=3	-0.279	-0.139
Sex*father=4	-0.103	0.068
Sex*father=5	0.537	0.589
Sex*Relat. bw parents=2	0.011	0.248
Sex*parents=3	-0.035	0.044
Sex*parents=4	0.158	-0.044
Sex*parents=5	0.075	0.388

Included sample: Coefficients obtained from the penalized logistic regression model using penalized maximum likelihood estimation (PMLE) in Pelotas dataset for those included in final analyses, after

exclusionary criteria. Available sample: Coefficients obtained from the penalized logistic regression model using PMLE in Pelotas dataset for all individuals meeting our inclusion criterion with no missing data on covariates (complete-case analysis). Relat. w/mother: adolescent's relationship with his(her) mother; Relat. w/father: adolescent's relationship with his(her) father; Relat. bw parents: relationship between parents (for all three, reference category 1="Great"). Interaction term is shown as "variable"*"variable".

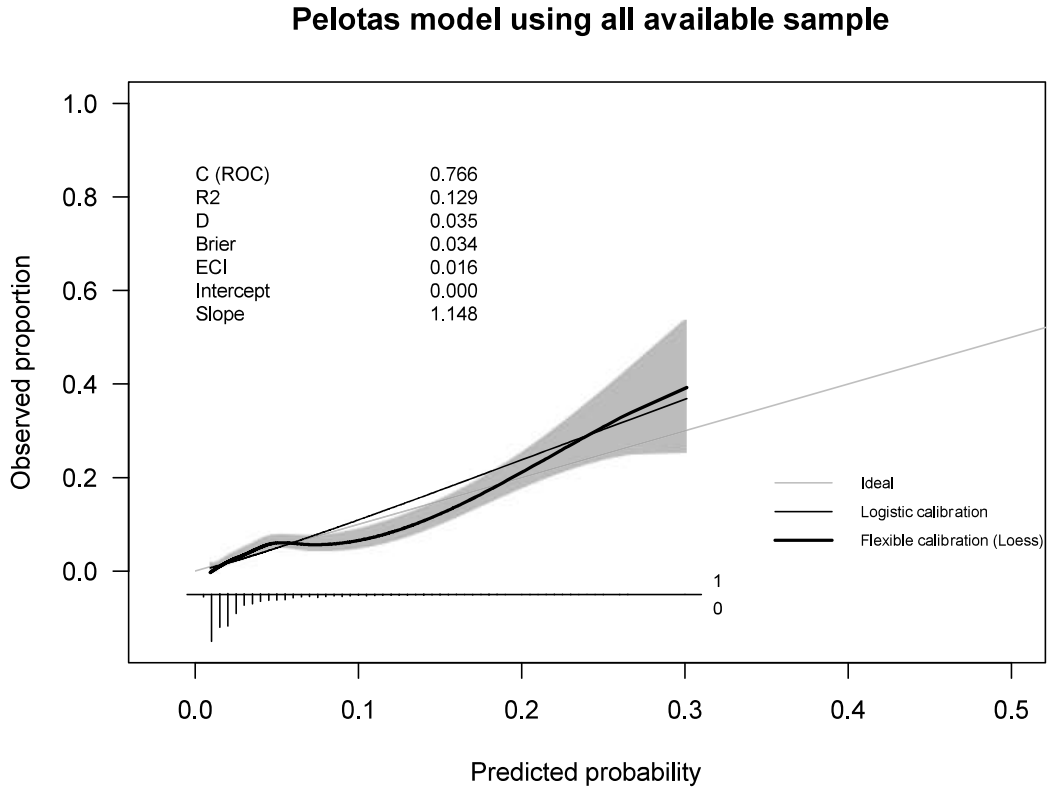


Figure S4. Predictive performance of the Pelotas model using all available sample
Calibration indicates the agreement between predicted probabilities and observed frequencies. A calibration plot is a graphical assessment of calibration, where the calculated predicted probability is shown on the x-axis, and observed frequency of the outcome on the y-axis. The flexible calibration line shows the result of a smoothing technique (Loess algorithm) used to estimate the observed probabilities of the binary outcome in relation to the predicted probabilities.
The grey shaded area indicates the 95% confidence interval around the estimated calibration line. The logistic calibration curve shows the result of the estimation of the observed proportions when a logistic model is used for the outcome as a function of the linear predictor of the developed model. The 45-degree line (labelled 'Ideal') has slope=1 and indicates perfect calibration. The distribution of predicted probabilities is shown at the bottom of the graphs. PMLE: Penalized maximum likelihood estimation; LR: Logistic regression. C (ROC): C-statistic; R2: Nagelkerke's R^2; D: Discrimination index; Brier: Brier score; ECI: Estimated Calibration Index; Intercept: Regression intercept of linear predictor; Slope: Regression slope of linear predictor.

Supplement 4. Exploratory analyses

In our study, supplementary strategies were implemented to further evaluate our model's properties in terms of predictive performance above and beyond currently established risk factors for depression; outcome specificity; and concentration of risk metrics.

We first assessed the relationship between our proposed model and two currently used strategies to identify adolescents at risk for developing depression: subsyndromal symptoms and family history of depression. Given our *a priori* decision not to rely on depressive symptoms as predictors, and only use variables directly obtained from the adolescent, we tested whether the inclusion of sub-threshold depressive symptoms and family history of depression, established risk factors for MDD, could provide additional predictive information to the model. Sub-threshold depressive symptoms were evaluated using the emotional sub-scale of the Strengths and Difficulties Questionnaire (SDQ) – parent-report at the 15-years assessment (no self-report measure was available for this age); and family history of depression was assessed using the maternal Self-Report Questionnaire (SRQ), only available for the 11-years assessment in Pelotas cohort study.

Net reclassification indices or improvement (NRI) have recently become popular statistics for measuring the prediction increment of new variables.[21] Measures of reclassification quantify the extent to which individuals are more appropriately classified into risk categories using a new model versus an old model. Individuals are placed into predefined risk categories based on their predicted absolute risk of experiencing the outcome (event) according to each model. Reclassification can be quantified using the NRI, that is the sum of 2 proportions: the proportion of events that move up through the risk categories upon using the new model; and the proportion of nonevents that move down through the risk categories upon using the new model. Current literature suggest the reporting of the "event NRI" (or NRI+) and "nonevent NRI" (NRI-) along with an overall NRI.

The impact of the inclusion of new variables into the Pelotas model for prediction of a depressive episode is shown in Table S7. Overall, the inclusion of these predictors failed to produce substantial improvements to the Pelotas model. The contribution of consolidated risk factors such as family history of depression (SRQ, either continuous or above the recommended threshold[22]) or sub-threshold depressive symptoms (emotional sub-scale of the SDQ – parent report) either alone or in combination, as shown in the first major row of the Table S7, produced only a small impact in all reclassification measures. These results suggest that the Pelotas model can provide predictive information to such an extent that the inclusion of further established risk factors could not produce meaningful classification improvement.

Additionally, we further assessed whether the developed risk score would behave differently according to either a positive depression screening or family history of depression status. By independently introducing each risk factor's main-effects and interaction terms with the Pelotas model linear predictor as covariates of the linear predictor itself for the depression outcome in Pelotas dataset, our results did not suggest a modification of predictive performance related to any of the factors' status (data available upon request).[23]

Conversely, we also evaluated whether the Pelotas model could provide an aid to decision-making strategies commonly used in current clinical settings. Using as baseline comparators three reduced models with consolidated risk factors for depressive disorder, namely biological sex, positive family history of depression and evidence of sub-threshold depressive symptoms, we aimed to assess if the use of Pelotas model could result in improvement above and beyond the information carried by these three models. As shown in the second major row, the added value of the Pelotas model to the baseline models was relevant in all measures. Evidence of greater and significant impact for the NRI- when compared to the NRI+ suggests its potential usefulness in reducing false-positive findings.

As a second step, we also evaluated the prediction performance of the Pelotas model linear predictor for all evaluated diagnoses at the 18-19 years' assessment in the Pelotas cohort, as can be seen in Table S8. The composite score was able to predict, to a lesser extent in comparison to the prediction of MDD, all other diagnostic categories assessed in the 1993 Pelotas Cohort.

Finally, as a third exploratory approach, we assessed the usefulness of the set of selected factors capacity itself in parsing high and low risk individuals across settings, after refitting the models in each dataset, using complementary metrics potentially useful in clinical decision-making (concentration of risk (CR) – the proportion of cases in the highest tenth of risk;[24] and high/low tenths ratio (HLTR) – the ratio between the proportion of cases in the highest and in the lowest tenths of the predicted risk range). In this conceptual validation, the Pelotas model's parsing capacity was high, with a CR of 31.9% and an incalculable HLTR (31.9/0), as its lowest tenth included only non-depressed participants. Similarly, for the E-Risk model, the CR was 18.3%, with a HLTR of 3.7, and the Dunedin model obtained a CR of 21.8%, with a HLTR of 3.9.

Table S7. Net Reclassification Improvement (NRI) analyses

		NRI	NRI+	NRI-	ΔC-statistic
Impact of predictor inclusion into Pelotas model	**Continuous SRQ**	0.041 (-0.004 – 0.096)	0.043 (0.000 – 0.100)	-0.002 (-0.011 – 0.006)	0.001
	Categorical SRQ	0.005 (-0.050 – 0.062)	0.000 (-0.056 – 0.057)	0.005 (-0.003 – 0.013)	-0.001
	Sub-threshold symptoms	0.037 (-0.038 – 0.121)	0.029 (-0.050 – 0.112)	0.008 (-0.004 – 0.021)	0.019
	Continuous SRQ + Sub-threshold symptoms	0.036 (-0.040 – 0.119)	0.029 (-0.044 – 0.110)	0.007 (-0.006 – 0.021)	0.021
Impact of Pelotas model inclusion into reduced models	**Reduced model 1**	0.222 (0.085 – 0.348)	0.043 (-0.092 – 0.171)	0.179 (0.155 – 0.202)	0.135
	Reduced model 2	0.200 (0.068 – 0.328)	-0.058 (-0.188 – 0.069)	0.258 (0.234 – 0.279)	0.145
	Reduced model 3	0.199 (0.051 – 0.342)	0.116 (-0.031 – 0.258)	0.083 (0.059 – 0.107)	0.142

Classification improvement according to the NRI (95% CI) at event rate.[21] NRI+: NRI for events (positive outcomes); NRI-: NRI for non-events (negative outcomes); ΔC-statistic: Change in C-statistic. SRQ: Self-Report Questionnaire; Categorical SRQ: SRQ>=8; Sub-threshold symptoms: According to emotional sub-scale of the Strengths and Difficulties Questionnaire (SDQ) – parent-report at the 15 years assessment in Pelotas.
Reduced model 1: sex and continuous SRQ; Reduced model 2: sex and categorical SRQ; Reduced model 3: sex, continuous SRQ and sub-threshold depressive symptoms.
Positive values indicate improvement and negative values indicate worsening in classification.

Table S8. Outcome specificity assessment of Pelotas model in the development sample

	R^2	C-Statistic	Calibration intercept	Calibration slope	Brier score	Emax
ADHD	0.05	0.70	-0.36	0.82	0.02	0.26
BD	0.03	0.66	-0.87	0.71	0.01	0.49
GAD	0.06	0.69	0.78	0.80	0.06	0.07
MDD	0.12	0.78	0.00	1.25	0.03	0.19
SAD	0.03	0.64	0.55	0.58	0.05	0.31

Prediction performance of the Pelotas model linear predictor for all evaluated diagnoses at the 18/19 years assessment in the Pelotas cohort. All psychiatric diagnoses were assessed by trained psychologists using an instrument derived from the Mini International Neuropsychiatric Interview.[8]
R^2: Nagelkerke's R^2; C-statistic: Concordance statistic, or area under the curve of the receiver operating characteristic (AUC-ROC); Calibration intercept: relates to calibration-in-the-large, which compares the mean of all predicted risks with the mean observed risk; Calibration slope: measure of agreement between observed and predicted risk of the event (outcome) across the whole range of predicted values; Brier score: Quadratic scoring rule that combines calibration and discrimination; Emax: Maximum absolute error in predicted probabilities. ADHD: Attention-deficit/hyperactivity disorder; BD: Bipolar disorder; GAD: Generalized anxiety disorder; MDD: Major depressive disorder; SAD: Social anxiety disorder.

References

1. Victora CG, Araújo CL, Menezes AM, et al. Methodological aspects of the 1993 Pelotas (Brazil) Birth Cohort Study. *Rev Saude Publica.* 2006;40(1):39-46.
2. Victora CG, Hallal PC, Araújo CL, Menezes AM, Wells JC, Barros FC. Cohort profile: the 1993 Pelotas (Brazil) birth cohort study. *Int J Epidemiol.* 2008;37(4):704-709.
3. Moffitt TE, E-Risk Study Team. Teen-aged mothers in contemporary Britain. *J Child Psychol Psychiatry.* 2002;43(6):727-742.
4. Odgers CL, Caspi A, Russell MA, Sampson RJ, Arseneault L, Moffitt TE. Supportive parenting mediates neighborhood socioeconomic disparities in children's antisocial behavior from ages 5 to 12. *Dev Psychopathol*. 2012;24(3):705-721.
5. Poulton R, Moffitt TE, Silva PA. The Dunedin Multidisciplinary Health and Development Study: overview of the first 40 years, with an eye to the future. *Soc Psychiatry Psychiatr Epidemiol.* 2015;50(5):679-693.
6. Rocha TB, Hutz MH, Salatino-Oliveira A, et al. Gene-environment interaction in youth depression: replication of the 5-HTTLPR moderation in a diverse setting. *Am J Psychiatry.* 2015;172(10):978-985.
7. Matthews T, Danese A, Wertz J, et al. Social isolation and mental health at primary and secondary school entry: a longitudinal cohort study. *J Am Acad Child Adolesc Psychiatry.* 2015;54(3):225-232.
8. Amorim P. Mini International Neuropsychiatric Interview (MINI): validação de entrevista breve para diagnóstico de transtornos mentais. *Revista Brasileira de Psiquiatria.* 2000;22:106-115.
9. Harrell F. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* 2nd ed. Cham, Switzerland: Springer; 2015. doi:10.1007/978-3-319-19425-7.
10. Moons KG, Donders AR, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol.* 2004;57(12):1262-1270.
11. Chaibub Neto E, Bare JC, Margolin AA. Simulation studies as designed experiments: the comparison of penalized regression models in the "large p, small n" setting. *PLoS One.* 2014;9(10):e107957.
12. Zou, H; Hastie, T. Regularization and variable selection via the elastic net.*J. R. Stat. Soc. Ser. B Stat. Methodol.* 67 (2005), no. 2, 301–320.
13. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-73.
14. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med.* 2015;162(1):55-63.
15. Thapar A, Collishaw S, Pine DS, Thapar AK. Depression in adolescence. *Lancet.* 2012;379(9820):1056-1067.
16. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35(29):1925-1931.
17. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res.* 2017;26(2):796-808.
18. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol.* 2015;68(3):279-289.
19. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol.* 2010;172(8):971-980.

20. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating.* New York, NY: Springer; 2009.
21. Pencina MJ, Steyerberg EW, D'Agostino RB. Net reclassification index at event rate: properties and relationships. *Stat Med.* 2017;36(28):4455-4467.
22. Mari JJ, Williams P. A validity study of a psychiatric screening questionnaire (SRQ-20) in primary care in the city of Sao Paulo. *Br J Psychiatry.* 1986;148:23-26.
23. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol.* 2016;69:245-247.
24. Kessler RC, Warner CH, Ivany C, et al. Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army Study To Assess Risk and rEsilience in Service members (Army STARRS). *JAMA Psychiatry.* 2015;72(1):49-57.
25. Bernardini F, Attademo L, Cleary SD, et al. Risk Prediction Models in Psychiatry: Toward a New Frontier for the Prevention of Mental Illnesses. *J Clin Psychiatry*. 2017;78(5):572-83.